

# Predicción de precios de acciones mediante modelos de aprendizaje automático.

Yunio Seijo Manso

A thesis submitted in conformity with the requirements for the MSc in Economics,  
Finance and Computer Science University of Huelva & International University of  
Andalusia



December 2022



Predicción de precios de acciones mediante modelos de  
aprendizaje automático.

Yunio Seijo Manso  
Máster en Economía, Finanzas y Computación  
Supervisor  
Manuel Emilio Gegúndez Arias  
Diego Marín Santos  
Universidad de Huelva y Universidad Internacional de Andalucía  
2022

## **Abstract**

In this master's thesis we develop a set of models for stock price forecasting using machine learning techniques on time series data.

In recent years, forecasting the dynamics of financial markets has been the subject of economic research. Predicting stock market prices is considered a monumental task and has garnered considerable attention, this task is quite difficult due to non-stationary, non-linear and chaotic data. Artificial neural networks (ANNs) are models that are currently used in a wide variety of fields and finance is no exception. In this thesis, random forest models, linear regression and neural networks with different architectures are used to predict stock prices. The data provided by the TradingView platform in daily temporality of five assets were used and the results of all models were compared using different performance metrics. The results showed that linear regression models and basic prediction models have better performance than models based on machine learning.

JEL classification: C02; C32; C45; C53; C63.

Keywords: Time series, artificial neural networks, machine learning, linear regression, financial markets.

## Resumen

En el presente trabajo fin de máster se desarrolla un conjunto de modelos para la predicción de precios de acciones usando técnicas de aprendizaje automático sobre datos en forma de una serie temporal.

En los últimos años, la previsión de la dinámica de los mercados financieros ha sido objeto de investigación económica. La predicción de los precios de los mercados bursátiles se considera una tarea monumental y ha acaparado una atención considerable, esta tarea es bastante difícil debido a los datos no estacionarios, no lineales y caóticos. Las redes neuronales artificiales (RNA) son modelos que se utilizan actualmente en una gran variedad de campos y las finanzas no es la excepción. En esta investigación, para predecir los precios de las acciones se utilizan modelos de bosques aleatorios, regresión lineal y redes neuronales con diferentes arquitecturas. Se utilizaron los datos provistos por la plataforma TradingView en temporalidad diaria de cinco activos y se compararon los resultados de todos los modelos utilizando diferentes métricas de rendimiento. Se obtuvo como resultados que los modelos de regresión lineal y modelos básicos de predicción tienen un mejor rendimiento que los modelos basados en aprendizaje automático.

**Palabras-Clave:** Series temporales, redes neuronales artificiales, aprendizaje automático, regresión lineal, mercados financieros.

## Acknowledgments

A mi familia, sobre todo a mi Madre y mi esposa, gracias por el apoyo incondicional que me han dado siempre, por no soltarme nunca de la mano y por confiar en mí. Soy quien soy gracias a ustedes. No olviden nunca que son mi motor. A la gran familia que me ha dado esta etapa académica, a los amigos que ya se han leído este documento varias veces, gracias por formar parte de mi vida, conseguiremos todos los sueños que nos trajeron hasta aquí hoy. Gracias a la Universidad Internacional de Andalucía por la oportunidad y la beca concedida, ha sido un lindo camino, a mis tutores por aceptar esta propuesta que sale de su zona habitual de investigación. Gracias a todos.

## Índice de contenidos

Capítulo 1. Propuesta de trabajo.....	1
1.1 Motivación.....	1
1.2 Objetivos generales y específicos.....	2
1.3 Propuesta metodológica .....	3
1.4 Estructura del documento .....	4
Capítulo 2. Introducción y estado del Arte.....	5
2.1 Introducción .....	5
2.2 Estado del arte.....	6
Capítulo 3. Marco teórico .....	12
3.2.1 Técnicas simples de predicción .....	13
3.2.2 Random Forest (Modelo de bosque aleatorio).....	13
3.2.3 Regresión lineal .....	13
3.2.4 Redes neuronales feedforward. ....	14
3.2.5 Redes neuronales convolucionales .....	15
3.2.6 Red recurrente LSTM.....	16
3.2.7 Redes recurrentes Bidireccional .....	16
Capítulo 4. Materiales .....	18
4.1 Bases de datos .....	18
4.2 Análisis exploratorio y estadístico de los datos.....	19
4.3 Conjuntos de análisis .....	27
4.4 Descripción de cada variable .....	29
Capítulo 5. Experimentación .....	32
5.1 Métricas de rendimiento para la evaluación de modelos de regresión .....	32
5.2 Modelos de referencias.....	33
5.3 Configuración de los modelos y resultados.....	34
5.4 Comparación de los resultados de los modelos.....	58
5.5 Análisis y discusión .....	61
Conclusiones.....	62
Referencias .....	63
Anexo 1.....	66

# Capítulo 1. Propuesta de trabajo

## 1.1 Motivación

La predicción de precios de acciones ha sido motivo de estudios desde diferentes enfoques, cuando se trata de los mercados de valores, además de su complejidad y dinamismo inherentes, ha habido un debate constante sobre la previsibilidad de los rendimientos de las acciones. Fama (1970) introdujo la hipótesis del mercado eficiente, según la cual el precio actual de un activo refleja siempre toda la información previa disponible de forma instantánea. También existe la hipótesis de la marcha aleatoria (Malkiel, 1999) que afirma que el precio de una acción cambia independientemente de su historia, es decir, que el precio de mañana sólo dependerá de información de mañana, independientemente del precio de hoy. Estas dos hipótesis determinan que no hay medios para predecir con exactitud el precio de una acción. Además, se han realizado una serie de experimentos que demuestran que una estrategia aleatoria puede superar a algunos de los métodos más clásicos de trading técnico como la Divergencia de Convergencia de la Media Móvil (MACD) y el Índice de Fuerza Relativa (RSI) (Biondo, Pluchino, Rapisarda, & Helbing, 2013).

Por otro lado, hay otros autores que afirman que, los precios de las acciones pueden predecirse al menos en cierta medida (Lo & MacKinlay, 2011) y una variedad de métodos para predecir y modelar el comportamiento de las acciones han sido objeto de estudio de muchas disciplinas, como la economía, la estadística, la física y la informática. Cabe mencionar que, en 2012, se estimó que aproximadamente el 85% de las operaciones en los mercados de valores fueron realizadas por algoritmos (Glantz & Kissell, 2013).

Las contradicciones entre diferentes autores y enfoques referentes a la predicción de precios de acciones en el mercado de valores, ha motivado esta investigación a realizar experimentos utilizando técnicas recientes de Inteligencia Artificial (IA). Una de las técnicas actualmente utilizadas son las redes neuronales, este trabajo se centra en comprobar su efectividad prediciendo el precio de algunas acciones y materias primas como el Oro.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

## **1.2 Objetivos generales y específicos**

### **Objetivo general**

Predecir el precio de cierre (close) para el día siguiente del precio del Oro y cuatro acciones que cotizan en la Bolsa de Nueva York (NYSE, del inglés), que forman parte del índice S&P500, aplicando técnicas de aprendizaje automático e identificando cuál presenta mejor desempeño basados en diferentes métricas.

### **Objetivos específicos**

- Obtener datos sobre los precios de cierre de acciones y materias primas junto con otras variables utilizadas en el análisis técnico.
- Procesar los datos para eliminar valores perdidos y analizar la relación entre las variables.
- Crear un modelo simple de predicción que sirva como base para la comparación con los modelos de aprendizaje automático.
- Realizar experimentos con distintas técnicas de aprendizaje automático y diferentes arquitecturas de redes neuronales artificiales.
- Evaluar los modelos comparándolos por diferentes métricas de rendimientos

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

### 1.3 Propuesta metodológica

Implementar diferentes técnicas de predicción sobre datos históricos de precios de acciones en temporalidad diaria, procesando y filtrando inicialmente los datos obtenidos desde la plataforma TradingView, luego hacer una selección de características con un enfoque empírico y basado en métodos específicos de selección. Aplicar estrategias muy simples, como predecir que el precio de cierre del próximo día será igual al anterior, este sería un modelo básico, hasta técnicas más complejas como modelos de redes neuronales recurrentes y compararlos midiendo sus rendimientos. Los datos de cada serie se dividirán en tres partes: el 80% de los datos comenzando en 1980 se utilizan para entrenar los modelos, el 19% por ciento de los datos se utiliza para validar todos los modelos y un 1% para pruebas. Se usará como lenguaje de programación Python, este posee bibliotecas como Pandas, Numpy, Sklearn, Keras y TensorFlow que facilitan el trabajo con el procesamiento de los datos y de implementación de los algoritmos de aprendizaje automático. En la Figura 1 se muestra el flujo de actividades descritas anteriormente.

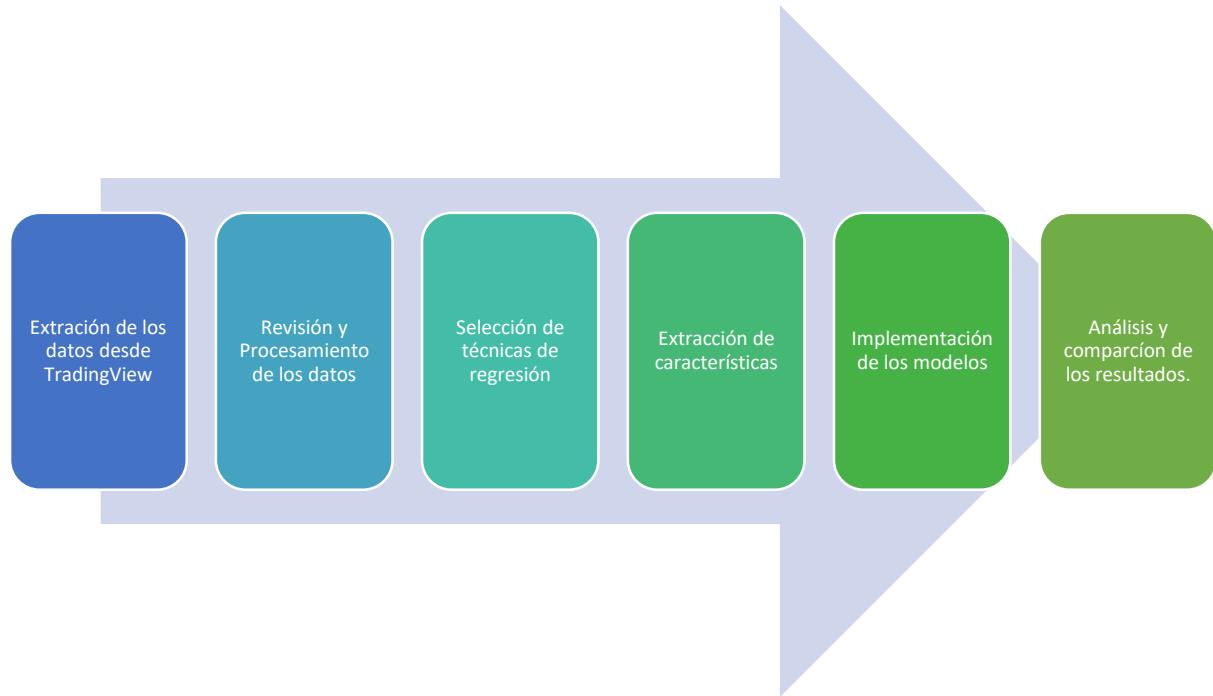


Figura 1. Flujo de actividades. Fuente: elaboración propia

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

## 1.4 Estructura del documento

El documento está dividido en 7 partes

- Propuesta del TFM: Se describen las principales motivaciones y los objetivos que se persiguen en la investigación.
- Introducción y estado del arte: Se abordan las características de las series temporales, y se comentan algunos estudios realizados sobre predicción en bolsa, a nivel mundial.
- Marco teórico: Se explican las técnicas usadas en el trabajo.
- Materiales: Se describen los datos utilizados, la temporalidad elegida, su procesamiento y tratamiento.
- Experimentación: Se explican los modelos básicos de referencia, se muestran las configuraciones de los modelos de redes neuronales, se comparan, analizan y discuten los resultados.
- Conclusiones. Se discuten ideas para próximas investigaciones y las conclusiones de este trabajo.
- Referencias y bibliografías: se incluyen las fuentes y recursos bibliográficos que se utilizaron como sustento teórico junto con el material consultado para hacer la investigación.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

## Capítulo 2. Introducción y estado del Arte

### 2.1 Introducción

La inteligencia artificial (IA) en las finanzas se vuelve un tema cada vez más popular tanto para la academia como para el sector de las finanzas. Numerosos estudios han sido publicados dando como resultado varios modelos e interpretaciones de cómo hacer predicciones. Dentro del campo del aprendizaje automático conocido como Machine Learning (ML) del inglés, y el aprendizaje profundo (*Deep Learnig*) ha recibido mucha atención recientemente, principalmente debido a su amplia divulgación, facilidades a lo hora de implementar estas técnicas que están disponible en varios lenguajes de programación, su rendimiento superior al de los modelos clásicos, así como el acceso a potentes sistemas de cómputos.

En los últimos 60 años, el campo de la IA ha experimentado su cuota de éxitos y fracasos. En la actualidad, los gobiernos de todo el mundo están compitiendo para crear instalaciones e investigaciones de IA superiores con vistas a que sea una palanca para un mayor poder económico e influencia. Entre 2012 y 2016, Estados Unidos invirtió 18.200 millones de dólares, frente a los 2.600 millones de China y los 850 millones del Reino Unido (Thornhill, 2018). El Fondo de Inversión de Pensiones del Gobierno japonés (el mayor gestor de ahorros para la jubilación del mundo) está considerando la IA para sustituir en última instancia a los gestores de fondos humanos. En febrero de 2018, BlackRock anunció que crearía un laboratorio con 6,3 billones de dólares de activos bajo gestión, la firma ya emplea el análisis de texto y analiza el tráfico de los sitios web corporativos y los datos de geolocalización de los teléfonos inteligentes, y ahora está estudiando en el ML para desplegarlo en la gestión de activos (Wigglesworth & Flood, 2018).

Los mercados financieros están evolucionando hacia un entorno mucho más tecnológico, en donde la ejecución de las órdenes muchas veces ya no es ejecutada por una persona, sino por un algoritmo (Aldridge, 2013). Según diferentes informes, entre los que se encuentra el estudio realizado por el instituto Alan Turin (Buchanan, 2019), se estima que entre el 60 y 73 % de las operaciones diarias son llevadas a cabo por algún tipo de algoritmo automatizado. Esto de alguna manera incrementa la liquidez del mercado y disminuye los costes de trading.

El mercado de valores representa un gran atractivo para los inversionistas, ya sea con el objetivo de proteger su patrimonio o incrementarlo, debido en gran parte, a especulaciones sobre el enorme potencial para la generación de ganancias que puede ofrecer. No obstante, el mercado presenta una alta variabilidad, la que puede generar fluctuaciones y cambios en los resultados esperados, creando escenarios donde la inversión no genera ganancias e inclusive se pierde parcial o totalmente el capital invertido. Este hecho es el resultado de diversos factores, tales como la oferta y demanda, la especulación, el rendimiento económico de la empresa, las políticas internacionales, etc.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

La historia y el presente de los métodos de aprendizaje automático aplicados a las finanzas y en especial a las series temporales económicas nos permiten que hoy podamos implementar varias técnicas de manera sencilla para predecir o al menos tener una aproximación de los precios de cierres en temporalidad diaria.

## 2.2 Estado del arte

### Breve historia sobre las redes neuronales

El primer modelo de una red neuronal artificial surge en 1943, de la mano de Warren McCulloch y Walter Harry Pitts. Pretende simular el funcionamiento de una neurona en el cerebro humano (McCulloch & Pitts, 1943). Es un modelo básico y que no se llegó a implementar físicamente debido a las limitaciones técnicas de la época. Hebb escribe el libro “The Organization of Behavior” donde postula que la información se representa en el cerebro mediante un conjunto de neuronas activas o inactivas y que el aprendizaje se localiza en las conexiones entre las neuronas (Hebb, 2005).

Rosenblatt (1958) elabora el modelo del perceptrón, modelo formado por una única neurona artificial que posee una salida binaria y que es capaz de variar sus pesos, pudiendo así resolver problemas de clasificación mediante una separación o regresión lineal. Sin embargo, el Perceptrón es incapaz de resolver problemas que no presentan una solución lineal, como sería el ejemplo de la puerta XOR. Marvin and Seymour (1969) publicaron en su libro “Perceptrons” las deficiencias de los modelos monocapa, lo que provoca una ralentización en el desarrollo tanto teórico como práctico de las redes neuronales hasta mediados de la década de los ochenta.

Con la publicación del algoritmo “Backpropagation” se produce el resurgir del desarrollo de las redes neuronales artificiales. Este algoritmo presenta una nueva forma de aprendizaje para las redes neuronales, mediante la retro propagación de errores desde las últimas capas y haciendo uso del descenso del gradiente para corregir este error (Rumelhart, Hinton, & Williams, 1986).

### Estudios para predecir el precio de acciones y materias primas

A continuación, se presentan brevemente algunos estudios donde se intenta predecir el precio de acciones y materias primas, así como su movimiento.

En lo que respecta a la inteligencia computacional, hay diferentes métodos para lograr predicciones precisas en el mercado de valores. Estos métodos van desde la computación evolutiva a través de algoritmos genéticos (Allen & Karjalainen, 1999) y el aprendizaje estadístico mediante algoritmos de *Support Vector Machines* (SVM) (Kim, 2003). Otros métodos incluyen redes neuronales, el modelado de componentes, el análisis textual basado en noticias (Melo, 2012), que propone un nuevo enfoque basado en la inteligencia colectiva. Trabajos más recientes utilizan las redes neuronales *Long-Short Term Memory* (LSTM) para predecir el movimiento del precio basado en un amplio rango de indicadores técnicos (Nelson, Pereira, & De Oliveira, 2017). Roondiwala, Patel, and Varma (2017) modelan y predicen los rendimientos de

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

las acciones del NIFTY 50 utilizando redes LSTM, recolectan cinco años de datos para usarlos en entrenamiento y validación del modelo, los datos históricos de las acciones. Aquí, los valores de entrenamiento se toman como los valores más recientes.

Los datos de prueba se toman de un 5% a un 10% del conjunto de datos. Eligen como características la fecha, apertura, máximo, mínimo, cierre y volumen, mantienen un tamaño de ventana de 22 días. Los datos abarcan desde enero de 2011 hasta diciembre de 2016. Para este experimento consideran la red neuronal recurrente y las redes de memoria a corto plazo, su modelo LSTM se compone de una capa de entrada secuencial seguida de dos capas LSTM y una capa densa con activación ReLU y finalmente una capa de salida densa con función de activación lineal. Utilizan Keras como *Frontend* y TensorFlow como *Backend* en su entorno de aprendizaje.

Para analizar la eficiencia del sistema utilizan el error cuadrático medio (RMSE). El error o la diferencia entre el objetivo y el valor de salida obtenido se minimiza utilizando el valor RMSE. En ese trabajo, en palabras del autor han utilizado una de las tecnologías de previsión más precisas utilizando la unidad de Redes Neuronales Recurrentes y Memoria a Corto Plazo que ayuda a los inversores, analistas o cualquier persona interesada en invertir en el mercado de valores proporcionándoles un buen conocimiento de la futura situación del mercado de valores.

Nelson et al. (2017) estudiaron la aplicabilidad de redes neuronales recurrentes, en particular las redes LSTM en el problema de la predicción de los movimientos del mercado de valores. Evalúan su rendimiento en términos de precisión y otras métricas a través de experimentos sobre datos de la vida real y analizan si presentan algún tipo de ganancia en comparación con los algoritmos tradicionales de aprendizaje automático.

Recogen datos históricos de las cotizaciones de determinados valores en el formato de una serie temporal de velas (apertura, cierre, máximo, mínimo y volumen) con una granularidad de 15 minutos, los datos usados son desde 2008 a 2015 para los valores que forman parte del índice IBovespa de la bolsa BM&F Bovespa, se realiza una transformación log-return como medio de normalización, así como para estabilizar la media y la varianza a lo largo de la serie temporal. Además de los datos de precios, se genera un total de 175 valores para cada periodo, que pretenden representar un conjunto muy diverso de características de la acción, como el precio futuro, el volumen, la intensidad de la tendencia del movimiento actual, patrones gráficos, entre otros.

En su estudio tratan de determinar si el precio de una acción en particular será mayor o menor que el precio actual en 15 minutos en el futuro. Ese modelo se regenera y entrena cada día de negociación sobre los datos históricos de precios y se utiliza para realizar predicciones cada 15 minutos utilizando el mismo modelo y las mismas ponderaciones hasta el final del día. Como resultados: Se ejecutaron una serie de experimentos utilizando el modelo propuesto y se recopiló el promedio de los

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

resultados con los datos de precios para una selección de valores diferentes de la bolsa brasileña en 2014. Para evaluar el rendimiento de la red, utilizaron métricas en torno al rendimiento del algoritmo y los resultados financieros se recogieron y se compararon con las líneas de base seleccionadas. Las métricas fueron la exactitud, la precisión, entre otras.

En general, el modelo propuesto en este artículo supera a las líneas de base con pocas excepciones. Los resultados pueden considerarse muy prometedores ya que ha demostrado ser capaz de predecir bien en comparación con otros enfoques empleados actualmente en la literatura. Aunque la dimensión de entrada es muy grande, el algoritmo ha demostrado una capacidad aceptable para aprender de ella sin necesidad de ninguna técnica de reducción de la dimensión, como la selección de características.

McNally, Roche, and Caton (2018) intentaron predecir el precio del Bitcoin utilizando el aprendizaje automático usando modelos ARIMA, RNN y RNN LSTM. Con datos de Open (precio de apertura), High (valor más alto), Low (valor más bajo), Close (precio de cierre) (OHLC) de CoinDesk y la dificultad de la tasa hash. La variable independiente para este estudio es el precio de cierre del Bitcoin.

Estos autores utilizaron ventanas de datos desde los 5 hasta a los 100 días múltiplos de 5. En su investigación siguen la metodología de minería de datos CRISP.1. La motivación para CRISP.1 frente a la más tradicional KDD (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) gira en el entorno empresarial de la tarea de predicción. Para evaluar el rendimiento de los modelos, utilizan el error cuadrático medio (RMSE) del precio de cierre y además codifican el precio previsto en una variable categórica que reflejan el precio al alza, a la baja o sin cambios. Este último paso permite obtener métricas de rendimiento adicionales que serían útiles para un operador en la formación de una estrategia de negociación: exactitud de la clasificación, especificidad, sensibilidad y precisión.

Los resultados obtenidos en los datos de validación reflejan que todos los modelos se esforzaron por aprender eficazmente de los datos de entrenamiento, el modelo redujo el error por debajo del 1%. En los datos de validación, el LSTM obtuvo un error del 8,07%, mientras que el RNN obtuvo un error del 7,15%. La precisión del 50,25% y el 52,78% logrados por los modelos de redes neuronales es una mejora marginal respecto a las probabilidades que se tienen en una tarea de clasificación binaria (precio al alza o a la baja), es decir, el 50%. La RNN no sirvió de nada cuando se utilizó una longitud temporal superior a 50 días. Por el contrario, la LSTM funcionó mejor en el rango de 50 a 100 días. En conclusiones, los modelos RNN y la LSTM, son evidentemente eficaces para la predicción de Bitcoin, siendo la LSTM más capaz de reconocer las dependencias a largo plazo. Sin embargo, una tarea de alta varianza de esta naturaleza hace que sea difícil transpolar en resultados impresionantes de validación. En consecuencia, sigue siendo una tarea difícil. Hay una línea muy fina entre sobreajustar un modelo e impedir que aprenda lo suficiente. El dropout (abandono) es una característica valiosa para ayudar a mejorar esto. Sin embargo, a

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

pesar de utilizar la optimización bayesiana para optimizar la selección de los dropouts, no se pueden garantizar buenos resultados de validación. A pesar de que las métricas de sensibilidad, especificidad y precisión indicaban un buen rendimiento, el rendimiento real del ARIMA basado en el error fue significativamente peor que los modelos de redes neuronales. La LSTM superó a la RNN marginalmente, pero no significativamente. Sin embargo, el LSTM tarda considerablemente más tiempo para entrenar.

En otra investigación se explora la viabilidad de utilizar indicadores de análisis técnico y redes profundas para estimar los precios de las acciones individuales y para predecir la tendencia a corto plazo del precio. Los datos experimentales de este estudio provienen de la información pública de la *Taiwán Stock Exchange Corporation* (TSEC). La fecha de negociación utilizada es del 4 de enero de 2019 al 21 de octubre de 2019, primer cuatrimestre (Q1) - segundo cuatrimestre (Q2) se establece como datos de entrenamiento, y el tercer cuatrimestre (Q3) son datos de prueba. Las características incluyen el precio de apertura diario, el precio de cierre, el precio más alto, el precio más bajo, las subidas y bajadas, el volumen y la tendencia, se utilizan indicadores de análisis técnico en general. Utilizan un modelo de red profunda utilizando cuatro capas de LSTM con función de pérdida de entropía cruzada, y 20 días de ventana (Lee, Liao, Yeh, & Chang, 2020).

Sus resultados muestran que la precisión media de todos los indicadores es de 0,75. Entre todos los indicadores, el MACD obtuvo la mayor precisión de estimación. Este estudio obtuvo una precisión media del 75% en la estimación del precio del tercer trimestre de TWSE 0050. El análisis posterior muestra que no hay diferencias significativas en las estimaciones del precio de las acciones para otros cuatro indicadores (Lee et al., 2020).

Mehtab, Sen, and Dutta (2020) proponen una gama de modelos basados en el aprendizaje profundo de memoria a corto y largo plazo (LSTM), para predecir con exactitud el movimiento del precio de las acciones del NIFTY 50 en la NSE de la India. Construyen cuatro modelos, tres modelos LSTM que son univariantes y el último modelo es multivariante. Utilizando valores históricos del índice NIFTY 50 para el período del 29 de diciembre de 2014 hasta el 28 de diciembre de 2018, el período de prueba que se extendió desde 31 de diciembre de 2018 hasta el 31 de julio de 2020.

Como datos de entrada, tres modelos utilizaron los datos de las dos semanas anteriores para predecir los valores de apertura para la semana siguiente, mientras que un modelo sólo utilizó datos de una semana. Siguen el enfoque de la regresión para ello, utilizan la variable Open como variable de respuesta y las demás variables como predictores. Observan que la regresión multivariante, MARS y el bosque aleatorio han superado a todos los demás modelos en el coeficiente de correlación. Sin embargo, la relación más baja de RMSE con respecto a la media de los valores de apertura reales la obtuvieron la regresión multivariante y el bosque aleatorio. Concluyen que la regresión multivariante y la regresión de bosque aleatorio fueron los

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

modelos más precisos en cuanto a su precisión de previsión en la serie temporal del NIFTY 50 (Mehtab et al., 2020).

Ding and Qin (2020) diseñan un modelo de red profunda basado en LSTM para predecir simultáneamente el precio de apertura, el precio más bajo y el precio más alto de una acción al día siguiente según el precio histórico de la acción y otros datos de análisis técnicos. El modelo de red asociada se compara con LSTM y red neuronal recurrente profunda basada en LSTM, y se verifica la viabilidad del modelo mediante la comparación de la precisión de los tres modelos. Los datos experimentales son los datos históricos reales de tres conjuntos de datos, uno de los cuales es el índice compuesto de Shanghái 000001 y los otros dos son acciones de PetroChina (código de código 601857) en la bolsa de Shanghai y ZTE (código 000063) en la bolsa de Shenzhen. El índice compuesto de Shanghai tiene 6112 datos históricos; PetroChina tiene 2688 datos históricos y ZTE tiene 4930 datos históricos. Cada conjunto de datos se divide en un conjunto de entrenamiento y un conjunto de prueba en orden cronológico en una proporción de 4:1. Cada conjunto de datos tiene siete parámetros técnicos. Se utilizan estos como atributos de entrada básicos, y el precio de apertura, precio más bajo y precio más alto del día siguiente como valores de salida del modelo. Se utiliza el método de normalización mínimo-máximo.

Se compara con la red LSTM y la red neuronal profunda-recurrente basada en LSTM (DRNN). El precio más alto de las acciones del día siguiente fue entrenado y predicho respectivamente por LSTM DRNN y Associated Net. Se propone un modelo de red asociada multivalor de red neuronal profunda-recurrente basada en LSTM para predecir múltiples precios de una acción simultáneamente. La viabilidad y la precisión de la red asociada se verifican comparando el modelo con el modelo de red LSTM y el modelo de red neuronal LSTM recurrente profundo. Constatan que el error cuadrático medio de los tres modelos disminuye gradualmente con el aumento de los tiempos de entrenamiento, el error cuadrático medio de la red LSTM es menor que el de los otros dos modelos, pero en la fase de prueba, el LSTM tiene el peor efecto de predicción y la precisión media más baja. Esto se debe a que el LSTM se sobreajusta a medida que aumenta el número de entrenamientos. Sus experimentos muestran que la precisión media del modelo asociado no sólo es mejor que la de los otros dos modelos. Además, puede predecir múltiples valores simultáneamente, y la precisión media de cada valor predicho es superior al 95% (Ding & Qin, 2020).

Livieris, Pintelas, and Pintelas (2020) proponen un nuevo modelo de previsión que se basa en la idea principal de explotar las ventajas de las técnicas de aprendizaje profundo. Concretamente, el modelo propuesto predice el valor del precio del Oro explotando la capacidad de las capas convolucionales, y la eficacia de las capas LSTM para identificar las dependencias a corto y largo plazo. Además, el modelo propuesto tiene la capacidad de predecir la dirección del movimiento del precio del Oro (aumento o disminución) en el día siguiente. Los datos utilizados en esta investigación se refieren a los precios diarios del Oro en USD desde enero de 2014 hasta abril de 2018 obtenidos de Yahoo Finanzas, los datos se dividieron en un

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

conjunto de entrenamiento y un conjunto de prueba. El conjunto de entrenamiento está formado por los precios diarios del Oro desde enero de 2014 hasta diciembre de 2017 (4 años), el conjunto de pruebas contiene los precios diarios de enero de 2018 a abril de 2018 (4 meses), los datos de entrenamiento como los de prueba fueron transformados utilizando un logaritmo natural ( $\ln$ ) para homogeneizar la variabilidad y la estabilidad de los patrones y reducir la tendencia exponencial. Utilizan ventanas de 4,6 y 9 días respectivamente como datos de entradas para predecir el precio del día siguiente. Evalúan el rendimiento del modelo CNN-LSTM propuesto frente al de los modelos LSTM (con una y dos capas LSTM) y los modelos de aprendizaje automático más avanzados de algunos de los artículos que le anteceden. Se apoyan en las bibliotecas de keras y tensorflow usan 50 épocas de entrenamiento y el optimizador de Adam con batch size de 128.

Obtienen como resultado que los modelos CNN-LSTM1 y CNN-LSTM2 tienen el mejor rendimiento global, en relación con todos los valores del horizonte. En cuanto al problema de la predicción del precio del Oro, CNN-LSTM2 superó considerablemente a todos los modelos de predicción, presentando la menor puntuación de MAE y RMSE, seguido de CNN-LSTM1. En conclusiones, señalan que, aunque los modelos LSTM constituyen ampliamente aceptados y eficientes para las series temporales, su utilización junto con capas convolucionales adicionales proporciona un impulso significativo en el aumento del rendimiento de la previsión.

Valverde Nieto (2021) pretende diseñar un sistema capaz de predecir el precio de cotización exacto de las acciones de Apple, Google, Microsoft y Amazon utilizando redes neuronales LSTM abarcando este problema desde el punto de vista de la economía, utiliza los datos para cada día los precios máximos, mínimos, de apertura y de cierre de las acciones, en USD, el volumen de operaciones y se añade el precio de cierre ajustado. Desde el 1 de enero de 2010 hasta el 7 de julio de 2021, extraídos de Yahoo Finanzas un total de 2899 días distintos. Parte de un modelo base obtenido de Kaggle y con una serie de modificaciones se formulan cinco variantes del modelo inicial que mejoran la predicción de precio de cierre que es la variable objetivo. Como resultados de la predicción los precios estimados son siempre inferiores que los precios futuros reales: las estimaciones infravaloran los precios. También son pesimistas los rendimientos estimados que generarían las acciones. Sería deseable que los errores en la predicción tuvieran una distribución uniforme. Que no se cumpla esto es síntoma de que el modelo puede mejorarse.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

## Capítulo 3. Marco teórico

En esta sección se comenta la definición de serie temporal y se analizan cómo funcionan las diferentes técnicas utilizadas en este trabajo para la predicción de precios, tanto las simples como las de aprendizaje automático.

### 3.1 Serie temporal

Una serie temporal (o simplemente una serie) es una secuencia de N observaciones (datos) ordenadas y equidistantes cronológicamente sobre una característica (serie univariante o escalar) o sobre varias características (serie multivariante o vectorial) de una unidad observable en diferentes momentos (Mauricio, 2007). De esta forma, podemos encontrar series temporales en cualquier ámbito de la vida cotidiana.

Las series temporales se estudian en diferentes áreas. El objetivo de analizarlas es, por un lado, entender o modelar el mecanismo que da lugar a una serie observada y por otro predecir los futuros valores de la serie basándose en el pasado de dicha serie o en otras series o factores relacionados.

En la estadística clásica, la principal preocupación es el análisis de las series temporales. El análisis de series temporales implica desarrollar modelos que capturen o describan mejor una serie temporal observada para apreciar las causas subyacentes. Este campo de estudio busca el "por qué" de un conjunto de datos de series temporales y los motivos que influyen en su dinámica (Shumway, Stoffer, & Stoffer, 2000).

Esto suele implicar hacer suposiciones sobre la forma de los datos y descomponer la serie temporal en componentes de constitución. La calidad de un modelo descriptivo viene determinada por lo bien que describe todos los datos disponibles y la interpretación que proporciona para informar mejor. "El objetivo principal del análisis de series temporales es desarrollar modelos matemáticos que proporcionen descripciones plausibles a partir de los datos de la muestra" (Shumway et al., 2000).

El intento de predecir el futuro se llama extrapolación en la estadística clásica. Los campos más modernos se concentran en esta tarea y la denominan previsión de series temporales. La previsión consiste en atribuir modelos que se ajustan a los datos históricos y utilizarlos para predecir las observaciones futuras (Yrigoyen, 2003). Una diferencia importante entre el análisis y la predicción es que esta última el futuro no está disponible y no puede ser estimado a partir de lo que ya ha ocurrido. En el contexto del aprendizaje automático la estimación del futuro sólo puede hacerse dividiendo el conjunto de datos en entrenamiento y validación. Y el modelo intenta aprender los datos de entrenamiento. La evaluación del rendimiento en los datos de validación garantiza nuestra precisión en la predicción del futuro.

### 3.2 Técnicas de predicción

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

### 3.2.1 Técnicas simples de predicción

Mañana igual que ayer. Esta es la técnica más simple que se usa, con un enfoque ingenuo, se predice que, el precio de cierre para el día ( $t$ ) será igual al del día ( $t-1$ ), pensando que no habrá una gran diferencia entre un día y el siguiente, este nos sirve de modelo base para comparar el rendimiento de las siguientes técnicas.

Ayer más variación, otra técnica básica donde se predice que el precio de cierre para el día ( $t$ ) será igual a el precio de cierre del día ( $t-1$ ) más la diferencia entre los precios de cierre del día ( $t-1$ ) y ( $t-2$ ) pensando que, la tendencia se mantendrá de un día al siguiente y no debería cambiar significativamente.

### 3.2.2 Random Forest (Modelo de bosque aleatorio)

El modelo de bosque aleatorio fue desarrollado por Breiman (2001). La técnica se basa en modelos de árboles de decisión y también se conoce como árboles de clasificación y regresión generalizados (CART), (Breiman, 2001) y (Booth, Gerdin, & McGroarty, 2014) brindan una descripción técnica detallada de la metodología de bosques aleatorios. El modelo de bosque aleatorio aplica la técnica de agregación de arranque a los aprendices de árbol. El algoritmo de bosque aleatorio es extremadamente útil para pronosticar, además de permitir la clasificación de si una variable es más o menos importante a la hora de construir el árbol de decisión.

Yeh, Chi, and Lin (2014) encuentran que las técnicas de bosque aleatorio son muy robustas y permiten la presencia de valores atípicos y ruido en el conjunto de entrenamiento. Los investigadores de JPMorgan consideran que el bosque aleatorio se muestra prometedor para el comercio de instrumentos del mercado del tesoro de EE.UU. a 10 años. Medeiros, Vasconcelos, Veiga, and Zilberman (2021) reconocen la propiedad de no linealidad con la dinámica de inflación y comparan 16 métodos ML con modelos estadísticos de referencia. En general, los métodos de ML con un gran conjunto de variables brindan resultados superiores en comparación con los puntos de referencia univariados y los modelos factoriales y encuentran que los bosques aleatorios es el mejor modelo que indica un grado de no linealidad en la dinámica de la inflación.

### 3.2.3 Regresión lineal

La regresión lineal múltiple permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta ( $Y$ ) se determina a partir de un conjunto de variables independientes llamadas predictores ( $X_1, X_2, X_3\dots$ ). Es una extensión de la regresión lineal simple. Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella (esto último se debe que analizar con cautela para no malinterpretar causalidad) (Rodrigo, 2016).

Los modelos lineales múltiples siguen la siguiente ecuación:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni}) + e_i \quad (1)$$

"Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas." Andrejs Dunkels

$\beta_0$ : es la ordenada en el origen, el valor de la variable dependiente Y cuando todos los predictores son cero.

$\beta_i$ : es el efecto promedio que tiene el incremento en una unidad de la variable predictora  $X_i$  sobre la variable dependiente Y, manteniéndose constantes el resto de variables. Se conocen como coeficientes parciales de regresión.

$\epsilon_i$ : es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo.

En este caso se utilizan un conjunto de datos usualmente usados en el análisis técnico como variables independientes para predecir el precio de cierre que sería nuestra variable dependiente.

### 3.2.4 Redes neuronales feedforward.

Una DNN (del inglés *Deep Neural Network*, Red Neuronal Profunda) es un conjunto de neuronas organizadas en una secuencia de múltiples capas, donde las neuronas reciben como entrada las activaciones neuronales de la capa anterior y realizan un cálculo simple (por ejemplo, una suma ponderada de la entrada seguida de una activación no lineal). Las neuronas de la red implementan conjuntamente un complejo mapeo no lineal de la entrada a la salida. Este mapeo se aprende a partir de los datos adaptando los pesos de cada neurona mediante una técnica llamada retro propagación de errores (Rumelhart et al., 1986). También se conocen estas redes por el nombre de red neuronal feedforward.

"El objetivo de una red *feedforward* es aproximar alguna función  $f^*$ . Por ejemplo, para un clasificador,  $y = f^*$  asigna una entrada  $x$  a una categoría  $y$ . Una red *feedforward* define un mapeo  $y = f(x; \theta)$  y aprende el valor de los parámetros  $\theta$  que resulta en la mejor función aproximada posible" (Goodfellow, Bengio, & Courville, 2016).

Las redes neuronales *feedforward* se llaman redes porque a menudo se representan componiendo muchas funciones diferentes. Por ejemplo, podemos tener tres funciones  $f(x) = f(3)(f(2)(f(1)(x))$ . En este caso,  $f(1)$  se llama primera capa de la red,  $f(2)$  se llama segunda capa, y así sucesivamente. La longitud total de la cadena da la profundidad del modelo. El nombre de "aprendizaje profundo" surgió de esta terminología. La última capa de una red *feedforward* se llama capa de salida, como se puede observar en la Figura 2 (Goodfellow et al., 2016).

"Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas." Andrejs Dunkels

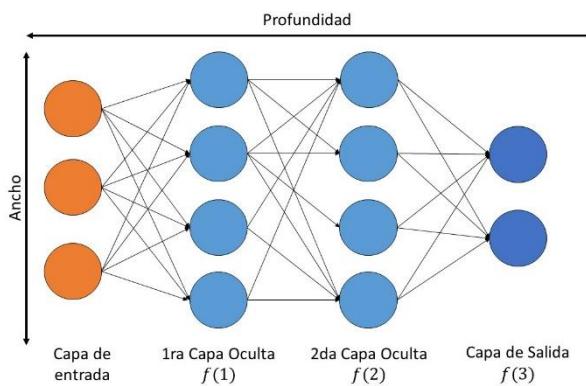


Figura 2. Red Neuronal *feedForward*. Fuente: elaboración propia.

"Cada elemento del vector puede interpretarse como un papel análogo al de una neurona. En lugar de pensar que la capa representa una única función vectorial también podemos pensar que la capa está formada por muchas unidades que actúan en paralelo, cada una representando una función de vector a escalar" (Goodfellow et al., 2016).

Las capas intermedias se denominan capas ocultas ya que los datos de entrenamiento no revelan la salida deseada para cada una de ellas. Por último, estas redes se denominan neuronales porque se inspiran libremente en la neurociencia. Las capas ocultas de la red suelen tener un valor vectorial. El tamaño de la dimensión de estas capas ocultas determina la amplitud del modelo.

### 3.2.5 Redes neuronales convolucionales

Las capas convolucionales y de agrupación (Rawat & Wang, 2017) son capas de pre procesamiento de datos especialmente diseñadas que tienen la tarea de filtrar los datos de entrada y extraer información útil que se utilizará como entrada normalmente en una capa de red totalmente conectada. Concretamente, las capas convolucionales aplican la operación de convolución entre los datos de entrada sin procesar y los núcleos de convolución que producen nuevos valores de características. Los datos de entrada deben tener forma de matriz estructurada, ya que esta técnica se originalmente destinado a la extracción de características de imágenes (Krizhevsky, Sutskever, & Hinton, 2017).

Una de las deficiencias de las arquitecturas totalmente conectadas es que la topología de la entrada es completamente ignorada. Las variables de entrada pueden presentarse en cualquier orden fijo sin que ello afecte al resultado del entrenamiento. Por el contrario, las imágenes o las representaciones espectrales del habla tienen una fuerte estructura local en dos dimensiones y las series temporales tienen una fuerte estructura en una dimensión (las variables o los píxeles), que están espacial o temporalmente cercanos y altamente correlacionados. Las correlaciones locales son la razón de las conocidas ventajas de extraer y combinar características locales antes de reconocer objetos espaciales o temporales. Las redes convolucionales fuerzan la extracción de características locales al restringir los campos receptivos de las unidades ocultas para que sean locales (LeCun & Bengio, 1995).

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

### 3.2.6 Red recurrente LSTM

Las RNN se entran mediante retro propagación (backpropagation) a través del tiempo (Werbos, 1990). Sin embargo, el aprendizaje de dependencias de largo alcance con RNN es difícil debido a los problemas de desaparición o explosión del gradiente (Bengio, Simard, & Frasconi, 1994; Hochreiter, Bengio, Frasconi, & Schmidhuber, 2001). La desaparición del gradiente en RNN se refiere a los problemas de que la norma del gradiente para los componentes a largo plazo disminuye exponencialmente rápido a cero, lo que limita la capacidad del modelo para aprender correlaciones temporales a largo plazo, mientras que la explosión del gradiente se refiere al evento opuesto. Para superar estos problemas (Hochreiter & Schmidhuber, 1997) introdujeron por primera vez la arquitectura de memoria a corto plazo (LSTM, ver Figura 3). Luego Gers, Schmidhuber, and Cummins (2000) incluyeron una celda de memoria y la mejoraron aún más con una puerta de olvido adicional. Ha sido la arquitectura de red recurrente más exitosa y recibió una gran popularidad en muchas aplicaciones posteriores.

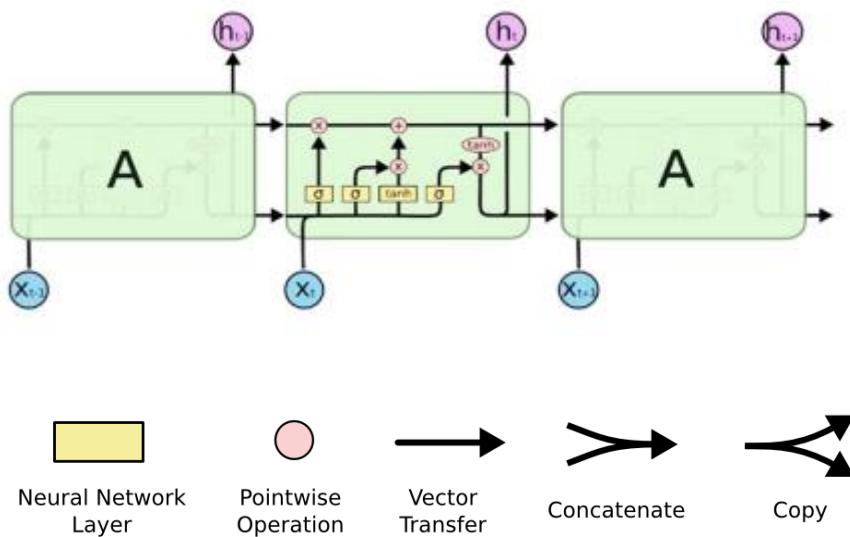


Figura 3. Red LSTM. Fuente: Olah (2015).

En el diagrama anterior, cada línea lleva un vector completo, desde la salida de un nodo hasta las entradas de los demás. Los círculos rosas representan operaciones puntuales, como la suma de vectores, mientras que los cuadros amarillos son capas de redes neuronales aprendidas. Las líneas que se fusionan denotan concatenación, mientras que una línea que se bifurca denota que su contenido se copia y las copias se dirigen a diferentes ubicaciones. La explicación detallada del funcionamiento de las redes LSTM guiado con diagramas se pueden ver en profundidad (Olah, 2015).

### 3.2.7 Redes recurrentes Bidireccional

Las LSTM bidireccionales propuestas por Schuster and Paliwal (1997) pueden entrenarse utilizando toda la información de entrada disponible en el pasado y en el futuro de un marco temporal específico, son una extensión de las LSTM tradicionales.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

El modelo incluye dos capas LSTM paralelas para proporcionar un bucle hacia delante y hacia atrás, como se ilustra en la Figura 4. La idea es que la red aproveche la información pasada y futura a través de las secuencias hacia delante y hacia atrás para hacer predicciones. En este caso, la información actual tiene como dependencia la información pasada y también está vinculada a la información futura. Las secuencias forward  $\rightarrow h$  y las secuencias hacia atrás  $\leftarrow h$ , respectivamente, están representadas por las flechas rojas y verdes de la Figura 4. Las redes neuronales recurrentes bidireccionales se pueden entrenar usando algoritmos similares a los de las redes neuronales recurrentes, porque las dos neuronas direccionales no tienen interacciones. Sin embargo, cuando se aplica la retro propagación a lo largo del tiempo, se necesitan procesos adicionales porque la actualización de las capas de entrada y salida no se puede realizar a la vez.

Los procedimientos generales para el entrenamiento son los siguientes: para el pase directo, primero se pasan los estados hacia delante y hacia atrás, luego se pasan las neuronas de salida. Para el paso hacia atrás, las neuronas de salida se pasan primero, luego los estados hacia adelante y los estados hacia atrás se pasan a continuación. Una vez realizadas las pasadas hacia adelante y hacia atrás, se actualizan los pesos (Schuster & Paliwal, 1997).

$$\begin{aligned}\vec{h}_t &= g\left(U_{\vec{h}}x_t + W_{\vec{h}}\vec{h}_{t-1} + b_{\vec{h}}\right) \\ \overleftarrow{h}_t &= g\left(U_{\overleftarrow{h}}x_t + W_{\overleftarrow{h}}\overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}\right) \\ y_t &= g\left(V_{\vec{h}}\vec{h}_t + V_{\overleftarrow{h}}\overleftarrow{h}_t + b_y\right)\end{aligned}\quad (2)$$

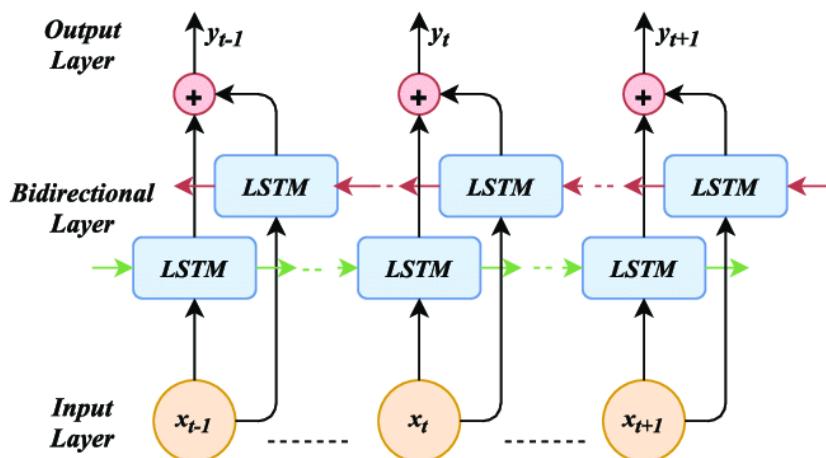


Figura 4 Estructura de una red Redes recurrentes Bidireccional. Fuente: Schuster and Paliwal (1997)

Althelaya, El-Alfy, and Mohammed (2018) y Sunny, Maswood, and Alharbi (2020) demostraron la superioridad de las redes neuronales bidireccionales sobre las redes LSTM en la predicción de precios de acciones.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

## Capítulo 4. Materiales

### 4.1 Bases de datos

Los datos de la investigación proceden de TradingView, que es una plataforma de gráficos y red social de más de 30 millones de traders e inversores de todo el mundo, donde se exploran las posibilidades de los mercados globales, es oficialmente el sitio web de inversión más popular del mundo. De esta plataforma se extrajeron datos de series del precio de cinco empresas y del precio del Oro. Las empresas fueron:

- Boing Company, cotiza en la bolsa de Nueva York con las siglas de BA.
- Walt Disney Company, cotiza en la bolsa de Nueva York con las siglas DIS.
- General Electric Company, cotiza en la bolsa de Nueva York con las siglas GE.
- International Business Machines, cotiza en la bolsa de Nueva York con las siglas IBM.

Cada una contiene los datos referentes al precio de cierre diarios en el formato de una serie temporal de velas (apertura, cierre, alto y bajo). Además, cuenta con un conjunto de indicadores técnicos, estos son cálculos matemáticos destinados a determinar o predecir características de las acciones en función de sus datos históricos. Se obtienen 22 variables para cada periodo de tiempo, las que se nombran a continuación, que pretenden representar o predecir un conjunto muy diverso de características de la acción, como precio futuro, volumen a ser negociado y la intensidad de la tendencia de movimiento actual. Estas variables que se usan en el análisis técnico serán descritas más adelante.

### Variables

Open, high, low, close, media de Bollinger, Banda Superior, Banda Inferior, 4EMA 3, 4EMA 6, 4EMA 13, 4EMA 21, Maximos, mitad, Minimos, Volume, Volume MA, Upper, Lower, ATR, Variacion %, Variacion, ema, ADX, sqzml, porcentaje a la media 1, RSI, Angulo a la media, stoch, diferencia, cambio y close\_log.

Las bases de datos contenían información con inicio en la década de los 70 y fin el día 10 de octubre del 2022. Se eliminaron todos los registros anteriores a 1980, pues había algunas variables con valores nulos o iguales a ceros. Se agregaron dos variables:

**Cambio**, variable que toma valor 1 si el precio fue superior al día anterior, -1 si fue menor y 0 en caso que fuera el mismo. Y **diferencia**, la cual expresa la primera diferencia del precio de cierre de la acción.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

Se realizó un preprocesamiento de los datos, siendo este primer paso de vital importancia para corregir múltiples deficiencias que podemos encontrar y para ser capaz de extraer realmente la información relevante para nuestro problema. Según distintas encuestas realizadas a científicos de datos, en torno a un 3% del tiempo de trabajo se centra en creación de conjuntos de datos de entrenamiento, 60% limpieza y organización de datos, 19% recopilación de conjuntos de datos, 9% patrones de minería de datos, 4% perfeccionamiento de algoritmos y 5% a otros (Press, 2016).

Problemas que usualmente tienen los datos y los científicos de datos deben tratar antes de comenzar a usarlos.

**Ausencia de valores:** Es muy común que, si estamos trabajando con un conjunto de datos grande, algunos valores estén vacíos. En el caso de la serie temporal de precios del Oro que se trata en esta investigación, la plataforma TradingView dispone de datos desde el año 1870.

Sin embargo, la mayoría de datos en temporalidad diaria, que es la temporalidad elegida en este trabajo eran valores perdidos, no es hasta 1980 que se puede notar que están disponibles todos los datos en temporalidad diaria.

**Inconsistencia de datos:** Ocurre en muchas ocasiones que cuando estamos procesando datos detectamos errores en el formato o en el tipo de alguno de ellos. En este caso no se detectan inconsistencias como pudieran ser valores negativos de precios.

## 4.2 Análisis exploratorio y estadístico de los datos

Para este análisis exploratorio de los datos se utilizó el módulo pandas-profiling. Es un módulo Open Source de Python con el que podemos hacer rápidamente un análisis exploratorio de datos con solo unas pocas líneas de código. Además, por si esto no fuera suficiente para convencernos de la utilidad de la herramienta, esta genera informes interactivos en formato web que pueden ser presentados a cualquier persona, aunque no sepa programar.

En resumen, lo que hace pandas profiling es ahorrarnos todo el trabajo de visualizar y entender la distribución de cada variable. Nos genera un informe con toda la información fácilmente visible. En esta sección se muestran algunos datos estadísticos de las variables usadas en el experimento por cada fuente analizada.

### Datos del Oro

Una vez visto una pequeña aproximación a los datos y ver que no tenemos datos perdidos o valores duplicados y sabemos qué tipo de variables vamos a utilizar seguimos con un análisis de la variable dependiente y vemos algunas interacciones con las demás variables.

### Análisis de las variables

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

En la Figura 5 se muestra el grado de correlación entre las variables utilizando el método de Pearson.

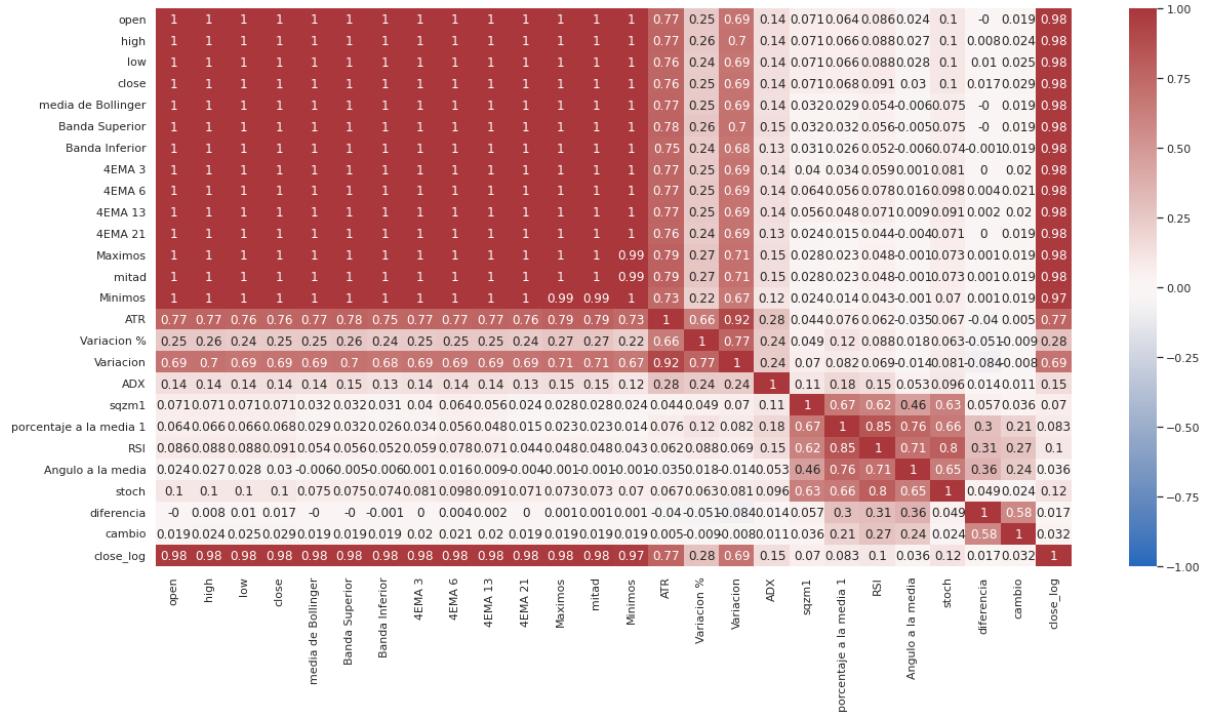


Figura 5. Matriz de correlación de Pearson para los datos del Oro. Fuente: elaboración propia.

En la Figura 5 el color rojo representa una alta correlación positiva, el blanco la ausencia de correlación y el azul alta correlación negativa. En esta matriz de correlación el número uno significa la mayor correlación posible, el cero, que no tiene correlación y menos uno la mayor correlación negativa posible.

## Oro

Tabla 1. Datos estadísticos de todas las variables del Oro. Fuente: elaboración propia a partir de Pandas. Fuente: elaboración propia.

	count	mean	std	min	25%	50%	75%	max
<b>open</b>	10715	749.394	506.112	252.25	364.85	435.3	1223.872	2063.468
<b>high</b>	10715	754.523	510.345	253.5	366	437.4	1232.095	2075.135
<b>low</b>	10715	744.491	501.956	252.1	363.45	432.8	1216.875	2034.7
<b>close</b>	10715	749.416	506.182	252.1	364.65	435	1224.05	2063.564
<b>media de Bollinger</b>	10715	748.315	505.124	254.871	364.257	434.005	1224.071	1972.063
<b>Banda Superior</b>	10715	767.808	517.532	256.558	371.881	446.338	1250.601	2053.162
<b>Banda Inferior</b>	10715	728.823	493.111	238.462	356.445	422.748	1200.515	1925.393
<b>4EMA 3</b>	10715	748.294	504.973	255.389	364.211	433.063	1224.107	1950.632
<b>4EMA 6</b>	10715	749.148	505.855	253.971	365.223	434.667	1224.067	2021.033
<b>4EMA 13</b>	10715	748.925	505.615	254.437	364.742	434.417	1224.272	1997.441
<b>4EMA 21</b>	10715	746.199	502.957	257.584	365.145	433.383	1219.311	1922.828

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>Maximos</b>	10715	789.69	530.522	261.7	383.7	461.9	1279.85	2075.135
<b>mitad</b>	10715	394.845	265.261	130.85	191.85	230.95	639.925	1037.568
<b>Minimos</b>	10715	707.139	477.512	252.1	349.3	410.5	1167.72	1878.25
<b>ATR</b>	10715	10.786	10.34	0.29	2.975	6.874	16.215	94.509
<b>Variacion (%)</b>	10715	1.152	1.105	0	0.404	0.978	1.584	15.287
<b>Variacion</b>	10715	10.032	12.096	0	1.5	6	15	160.4
<b>ADX</b>	10715	23.711	9.514	5.431	16.638	21.719	29.086	61.924
<b>sqzm1</b>	10715	0.812	24.814	- 151.337	-8.123	-0.079	9.38	162.871
<b>porcentaje a la media 1</b>	10715	0.174	3.135	-23.431	-1.434	-0.071	1.664	41.785
<b>RSI</b>	10715	50.522	13.03	11.031	41.416	49.643	59.513	93.689
<b>Angulo a la media</b>	10715	0.761	31.313	- 253.038	-15.143	0.627	16.785	297.475
<b>stoch</b>	10715	49.721	28.135	0	24.028	48.773	75.957	100
<b>diferencia</b>	10715	0.102	9.902	-143.5	-2.6	0.1	3.1	85.15
<b>cambio</b>	10715	0.026	0.992	-1	-1	1	1	1
<b>close_log</b>	10715	6.407	0.637	5.53	5.899	6.075	7.11	7.632

En la Figura 6 se representa la forma que sigue la serie de precios del Oro dividida en tres conjuntos, conjunto de entrenamiento representado con el color azul, conjunto de validación representado con el color verde y conjunto de prueba representado con el color naranja.



Figura 6. Precios de cierres del Oro. Fuente: elaboración propia

## Boing Company

Tabla 2. Datos estadísticos de todas las variables de Boing Company. Fuente: elaboración propia.

	count	mean	std	min	25%	50%	75%	max
<b>open</b>	10786	74.038	85.293	2.278	19.125	44.188	85.89	446.010

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>high</b>	10786	74.903	86.27	2.296	19.312	44.821	86.5	446.010
<b>low</b>	10786	73.135	84.261	2.222	18.943	43.562	85.072	440.190
<b>close</b>	10786	74.02	85.263	2.278	19.125	44.345	85.85	440.620
<b>media de Bollinger</b>	10786	73.902	85.124	2.332	19.055	44.025	85.081	421.171
<b>Banda Superior</b>	10786	77.543	89.682	2.413	19.778	46.119	88.372	447.412
<b>Banda Inferior</b>	10786	70.262	80.839	2.202	18.343	41.845	81.941	404.972
<b>4EMA 3</b>	10786	73.898	85.046	2.351	19.062	43.902	84.542	417.753
<b>4EMA 6</b>	10786	73.991	85.207	2.288	19.161	44.157	85.394	433.052
<b>4EMA 13</b>	10786	73.967	85.164	2.299	19.132	44.102	85.252	428.698
<b>4EMA 21</b>	10786	73.668	84.689	2.506	19.007	43.92	83.85	390.997
<b>Maximos</b>	10786	81.258	94.24	2.667	20.438	49.625	90.38	446.010
<b>mitad</b>	10786	40.629	47.12	1.333	10.219	24.812	45.19	223.005
<b>Minimos</b>	10786	65.999	76.005	2.222	17.688	39.05	76.4	361.525
<b>Volume</b>	10786	4487809	5610299	44996	2118345	3215924	4871464	103212800.000
<b>Volume MA</b>	10786	4483073	4948410	893276.8	2381007	3432867	4730855	70661250.000
<b>ATR</b>	10786	1.948	2.831	0.033	0.357	1.125	2.002	33.950
<b>Variacion %</b>	10786	2.4	1.537	0.272	1.429	2.007	2.908	25.440
<b>Variacion</b>	10786	1.768	2.707	0.019	0.312	0.94	1.86	39.990
<b>ADX</b>	10786	23.971	9.554	5.638	17.091	22.223	28.797	63.823
<b>sqzm1</b>	10786	0.174	7.019	-130.216	-0.734	0.133	1.491	51.317
<b>porcentaje a la media 1</b>	10786	0.472	5.66	-59.763	-2.421	0.703	3.709	49.254
<b>RSI</b>	10786	52.071	12.858	8.524	42.97	52.028	61.23	94.192
<b>Angulo a la media</b>	10786	2.182	59.123	-870.224	-26.3	3.712	32.633	557.675
<b>stoch</b>	10786	54.065	26.9	1.596	29.818	55.883	78.812	98.563
<b>diferencia</b>	10786	0.012	2.472	-41.93	-0.36	0	0.4	31.050
<b>cambio</b>	10786	0.005	0.981	-1	-1	0	1	1.000
<b>close_log</b>	10786	3.694	1.177	0.823	2.951	3.792	4.453	6.088

En la Figura 7 se representa la forma que sigue la serie de precios de *Boing Company* dividida en tres conjuntos, conjunto de entrenamiento representado con el color azul, conjunto de validación representado con el color verde y conjunto de prueba representado con el color naranja. Esta estructura de la serie precios con valores que

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

siguen una distribución diferente en el periodo de validación al de entrenamiento es un gran problema para los modelos de aprendizaje automático.

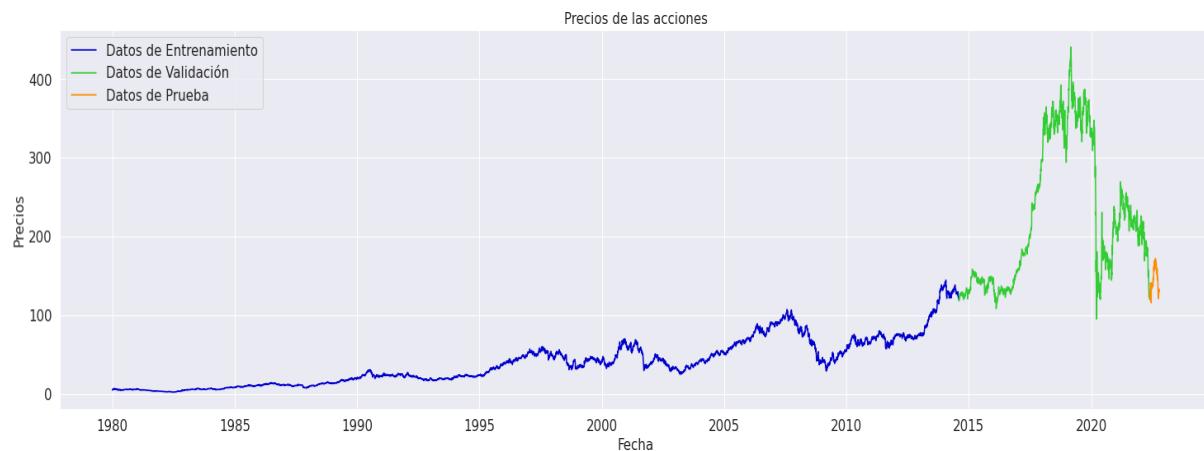


Figura 7. Precios de Boing Company. Fuente: elaboración propia

## International Business Machines (IBM)

Tabla 3. Datos estadísticos de todas las variables de IBM. Fuente: elaboración propia.

	count	mean	std	min	25%	50%	75%	max
<b>open</b>	10784	80.094	54.859	9.79	27.043	79.613	124.289	205.723
<b>high</b>	10784	80.849	55.268	9.97	27.282	80.42	125.463	206.220
<b>low</b>	10784	79.382	54.479	9.701	26.804	78.861	123.217	204.692
<b>close</b>	10784	80.121	54.882	9.79	27.043	79.637	124.294	206.124
<b>media de Bollinger</b>	10784	80.022	54.837	10.252	26.959	79.325	123.529	202.845
<b>Banda Superior</b>	10784	82.972	56.53	10.603	27.921	82.461	128.933	209.902
<b>Banda Inferior</b>	10784	77.071	53.228	9.823	25.964	76.946	119.817	200.289
<b>4EMA 3</b>	10784	80.02	54.808	10.222	26.969	79.198	123.485	201.463
<b>4EMA 6</b>	10784	80.096	54.86	10.003	27.014	79.602	124.039	203.941
<b>4EMA 13</b>	10784	80.076	54.845	10.074	27.02	79.477	123.909	203.009
<b>4EMA 21</b>	10784	79.84	54.708	10.506	27.003	79.832	123.6	198.066
<b>Maximos</b>	10784	85.765	58.049	11.074	29.192	85.888	134.143	206.220
<b>mitad</b>	10784	42.883	29.025	5.537	14.596	42.944	67.072	103.110
<b>Minimos</b>	10784	73.704	51.445	9.701	25.073	74.617	114.56	188.654
<b>Volume</b>	10784	68408 97	448740 4	4606	404988 0	578434 7	834108 0	72704600.00 0
<b>Volume MA</b>	10784	68380 65	318746 4	148740 9	465518 5	620807 3	835087 2	22649070.00 0
<b>ATR</b>	10784	1.603	1.295	0.125	0.473	1.331	2.357	11.150
<b>Variacion %</b>	10784	1.947	1.21	0	1.173	1.624	2.36	21.989

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>Variacion</b>	10784	1.468	1.305	0	0.448	1.137	2.054	17.432
<b>ADX</b>	10784	25.329	9.96	8.691	17.754	23.253	30.795	64.813
<b>sqzm1</b>	10784	0.147	3.773	-25.55	-1.113	0.183	1.58	17.116
<b>porcentaje a la media 1</b>	10784	0.309	4.503	-31.716	-2.238	0.413	2.964	22.352
<b>RSI</b>	10784	51.316	13.178	5.29	41.828	51.436	60.833	88.963
<b>Angulo a la media</b>	10784	1.379	44.643	-	-22.892	1.506	26.785	267.221
<b>stoch</b>	10784	52.752	28.049	1.966	26.475	54.521	79.003	98.937
<b>diferencia</b>	10784	0.01	1.485	-16.775	-0.418	0	0.478	12.895
<b>cambio</b>	10784	0.012	0.99	-1	-1	0	1	1.000
<b>close_log</b>	10784	4.071	0.858	2.281	3.297	4.377	4.823	5.328

En la Figura 8 se representa la forma que sigue la serie de precios de la compañía IBM dividida en tres conjuntos, conjunto de entrenamiento representado con el color azul, conjunto de validación representado con el color verde y conjunto de prueba representado con el color naranja.



Figura 8. Precios de cierres de IBM. Fuente: elaboración propia

## Walt Disney Company

Tabla 4. Datos estadísticos de todas las variables de Walt Disney Company. Fuente: elaboración propia.

	<b>count</b>	<b>mean</b>	<b>std</b>	<b>min</b>	<b>25%</b>	<b>50%</b>	<b>75%</b>	<b>max</b>
<b>open</b>	10785	39.757	43.501	0.855	9.473	25.08	42.424	200.185
<b>high</b>	10785	40.145	43.872	0.873	9.545	25.404	42.77	203.020
<b>low</b>	10785	39.337	43.068	0.842	9.369	24.855	42.1	195.400
<b>close</b>	10785	39.751	43.469	0.855	9.473	25.12	42.42	201.910

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>media de Bollinger</b>	10785	39.66	43.409	0.887	9.454	25.09	42.23	193.976
<b>Banda Superior</b>	10785	41.216	45.046	0.921	9.894	26.151	43.52	199.934
<b>Banda Inferior</b>	10785	38.103	41.825	0.839	9.113	23.988	41.138	189.127
<b>4EMA 3</b>	10785	39.658	43.385	0.886	9.46	25.079	42.222	192.345
<b>4EMA 6</b>	10785	39.729	43.448	0.876	9.484	25.107	42.243	195.989
<b>4EMA 13</b>	10785	39.71	43.432	0.881	9.469	25.125	42.293	194.978
<b>4EMA 21</b>	10785	39.488	43.226	0.847	9.42	24.879	41.472	185.218
<b>Maximos</b>	10785	42.777	46.769	0.946	10.221	27.002	44.34	203.020
<b>mitad</b>	10785	21.389	23.384	0.473	5.111	13.501	22.17	101.510
<b>Minimos</b>	10785	36.128	39.819	0.738	8.609	22.66	38.645	181.010
<b>Volume</b>	10785	7950544	6794508	331796	4388790	6555910	9522100	203492300.000
<b>Volume MA</b>	10785	7946753	4454934	1872005	5032102	7118562	9466670	55085200.000
<b>ATR</b>	10785	0.884	1.101	0.009	0.198	0.568	1.103	11.537
<b>Variacion %</b>	10785	2.207	1.424	0.197	1.3	1.837	2.703	22.283
<b>Variacion</b>	10785	0.808	1.063	0.003	0.166	0.499	0.998	14.120
<b>ADX</b>	10785	24.442	9.863	7.722	17.004	22.349	29.894	73.569
<b>sqzm1</b>	10785	0.113	2.525	-24.457	-0.376	0.051	0.681	16.875
<b>porcentaje a la media 1</b>	10785	0.683	5.017	-37.901	-1.911	0.922	3.611	28.422
<b>RSI</b>	10785	52.483	12.819	12.158	43.508	52.769	61.526	88.135
<b>Angulo a la media</b>	10785	2.935	50.321	- 500.903	-23.738	3.621	30.262	284.153
<b>stoch</b>	10785	54.526	26.99	1.173	30.237	57.15	79.156	98.642
<b>diferencia</b>	10785	0.009	1.003	-13.7	-0.187	0	0.2	21.030
<b>cambio</b>	10785	0.017	0.984	-1	-1	0	1	1.000
<b>close_log</b>	10785	2.939	1.431	-0.156	2.248	3.224	3.748	5.308

En la Figura 9 se representa la forma que sigue la serie de precios de DIS dividida en tres conjuntos, conjunto de entrenamiento representado con el color azul, conjunto de validación representado con el color verde y conjunto de prueba representado con el color naranja.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

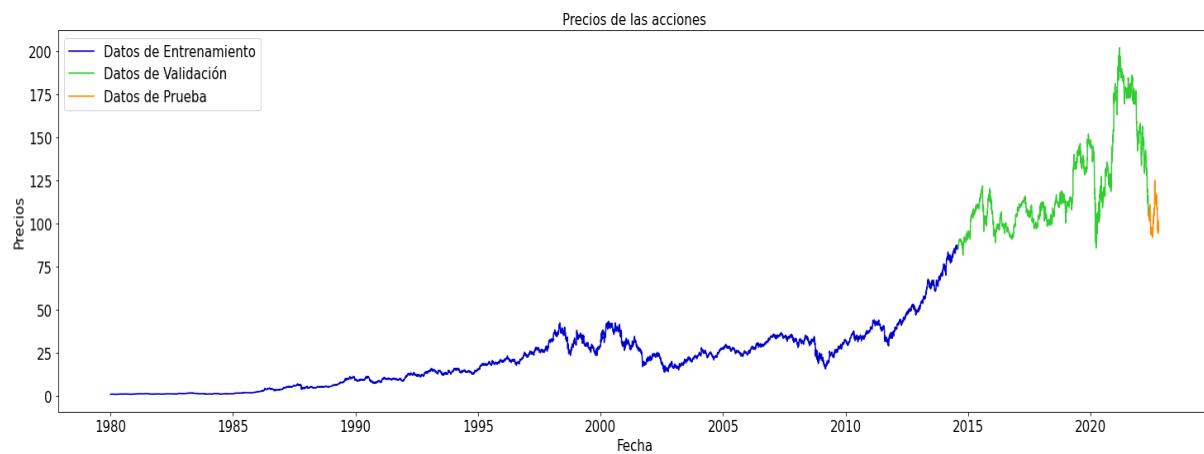


Figura 9. Precios de cierres de DIS. Fuente: elaboración propia

## General Electric Company

Tabla 5. Datos estadísticos de todas las variables General Electric Company. Fuente: elaboración propia.

	<b>count</b>	<b>mean</b>	<b>std</b>	<b>min</b>	<b>25%</b>	<b>50%</b>	<b>75%</b>	<b>max</b>
<b>open</b>	10784	134.41	104.059	7.169	41.25	107.652	218.936	458.004
<b>high</b>	10784	135.844	105.208	7.229	41.571	108.729	221.476	465.213
<b>low</b>	10784	132.886	102.87	7.049	41.01	106.196	216.074	457.043
<b>close</b>	10784	134.39	104.083	7.169	41.25	107.499	218.611	461.368
<b>media de Bollinger</b>	10784	134.337	103.998	7.444	41.056	106.968	219.096	450.383
<b>Banda Superior</b>	10784	139.763	108.109	7.624	42.465	111.489	227.176	467.625
<b>Banda Inferior</b>	10784	128.911	100.054	7.194	39.731	102.457	210.355	438.638
<b>4EMA 3</b>	10784	134.336	103.937	7.476	40.906	106.554	218.883	449.675
<b>4EMA 6</b>	10784	134.377	104.035	7.324	41.25	107.055	218.761	455.941
<b>4EMA 13</b>	10784	134.366	104.008	7.388	41.126	106.975	218.787	454.150
<b>4EMA 21</b>	10784	134.235	103.736	7.621	40.23	106.654	220.441	438.675
<b>Maximos</b>	10784	145.124	111.899	7.81	43.273	116.165	235.648	465.213
<b>mitad</b>	10784	72.562	55.95	3.905	21.636	58.082	117.824	232.606
<b>Minimos</b>	10784	122.289	95.262	7.049	37.926	96.733	200.355	415.231
<b>Volume</b>	10784	4330382	4659169	312	1715637	2719722	5243283	97913970.000
<b>Volume MA</b>	10784	4325504	3937192	747554.4	1847563	2694549	5581309	39190960.000
<b>ATR</b>	10784	3.172	3.26	0.084	0.759	2.462	4.064	26.576
<b>Variacion %</b>	10784	2.185	1.531	0	1.252	1.796	2.616	21.935
<b>Variacion</b>	10784	2.957	3.402	0	0.671	1.999	3.768	38.524
<b>ADX</b>	10784	24.053	9.178	7.594	16.974	22.45	29.402	58.352
<b>sqzm1</b>	10784	0.228	7.252	-37.845	-1.744	0.14	2.456	50.755

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>porcentaje a la media 1</b>	10784	0.286	4.983	-38.772	-2.018	0.557	3.079	22.743
<b>RSI</b>	10784	51.642	12.597	14.3	42.803	51.709	60.577	86.866
<b>Angulo a la media</b>	10784	1.413	50	-428.137	-22.409	1.138	26.948	399.432
<b>stoch</b>	10784	53.03	27.334	1.346	28.021	54.006	78.432	98.962
<b>diferencia</b>	10784	0.005	3.055	-36.14	-0.769	0	0.769	31.238
<b>cambio</b>	10784	-0.001	0.981	-1	-1	0	1	1.000
<b>close_log</b>	10784	4.469	1.057	1.97	3.72	4.677	5.387	6.134

En la Figura 10 se representa la forma que sigue la serie de precios de GE dividida en tres conjuntos, conjunto de entrenamiento representado con el color azul, conjunto de validación representado con el color verde y conjunto de prueba representada con el color naranja.

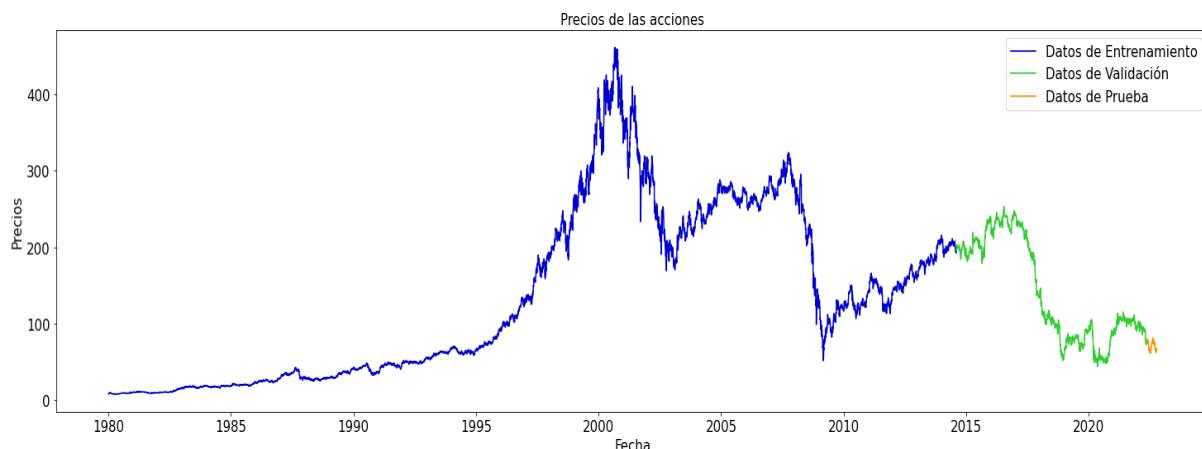


Figura 10. Precios de cierres del GE. Fuente: elaboración propia

### 4.3 Conjuntos de análisis

Los datos se dividen en tres conjuntos con el objetivo de tomar un conjunto para entrenar los modelos, otro para validar el aprendizaje de los algoritmos y otro para hacer predicciones de pruebas.

El conjunto de entrenamiento representa el 80.9% de los datos totales disponibles, como se trabaja con una serie temporal y el orden es importante pues serían el 80% inicial de datos de la serie. El conjunto de validación representa el 18.2% de los datos totales y el de prueba el 0.009.

Esta distribución de los datos nos aleja un poco de la realidad, pues se usarían datos para entrenar desde el 1980 hasta 2015 aproximadamente para entrenar los modelos

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

e intentar predecir lo que sucede tres años después basado en un entrenamiento muy alejado en el tiempo es un problema significativo. Se parte de cinco conjuntos de variables diferentes y se aplican todos los modelos de aprendizaje automático mencionados anteriormente.

### Primer conjunto de variables utilizadas

('close', 'Banda Superior', 'Banda Inferior', '4EMA 3', '4EMA 6', '4EMA 13', '4EMA 21', 'ATR', 'Variacion %', 'Variacion', 'ADX', 'sqzm1', 'porcentaje a la media 1', 'stoch', 'diferencia', 'cambio'). Usando como variables de entradas tres retrasos de cada una de ellas.

Por ejemplo: de la variable “diferencia” en el día (t), (t-1) y (t-2) sería la entrada para predecir el close (t+1), así todas las variables y sus tres retrasos serían usadas para predecir el close (t+1). Similar a como se muestra en la Figura 11.

	closetarget	close0	close1	close2	Banda Superior0	Banda Superior1	Banda Superior2	Banda Inferior0	Banda Inferior1	Banda Inferior2	...
time											
1980-01-06	602.5	627.0	603.599976	625.000000	589.394549	571.985458	556.623745	416.348301	415.805011	415.938155	...
1980-01-07	599.0	602.5	627.000000	603.599976	600.142062	589.394549	571.985458	420.696026	416.348301	415.805011	...
1980-01-08	598.5	599.0	602.500000	627.000000	608.559250	600.142062	589.394549	427.735981	420.696026	416.348301	...
1980-01-09	646.0	598.5	599.000000	602.500000	615.374714	608.559250	600.142062	436.396708	427.735981	420.696026	...
1980-01-10	671.0	646.0	598.500000	599.000000	628.502168	615.374714	608.559250	442.316873	436.396708	427.735981	...

Figura 11. Entrada y salida correspondiente de los datos. Fuente: elaboración propia.

### Segundo conjunto de variables utilizadas

Partiendo del conjunto uno donde se tienen todas las variables con tres retrasos cada una, se hace una selección de las 10 características más importantes utilizando el método de selección hacia adelante. Las variables seleccionadas fueron ('close0', 'Banda Superior0', '4EMA 210', '4EMA 211', '4EMA 212', 'ADX2', 'sqzm10', 'porcentaje a la media 10', 'diferencial', 'stoch2').

### Tercer conjunto de variables utilizadas

Partiendo del conjunto uno donde se tienen todas las variables con tres retrasos cada una, se hace una selección de las 10 características más importantes utilizando el método de selección eliminación hacia atrás. Las variables seleccionadas fueron ('close0' '4EMA 131' '4EMA 211' 'ATR0' 'ATR1' 'Variacion %0' 'Variacion2' 'sqzm10' 'porcentaje a la media 10' 'porcentaje a la media 11').

### Cuarto conjunto de variables utilizadas

Partiendo de todas las variables iniciales y con cinco retrasos de cada una se hace una selección de las 10 características más importantes utilizando el método de Cresta. Las variables seleccionadas fueron ('close0', '4EMA 60', 'ATR0', 'ATR1', 'ATR2', 'ADX1', 'sqzm12', 'sqzm13', 'diferencia0', 'cambio4').

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

## Quinto conjunto de variables utilizadas.

De las 16 variables seleccionadas utilizando métodos de filtrado, se hace una segunda selección de las 10 características más importantes utilizando el método de Cresta. Las variables seleccionadas fueron ('close0', 'close1', '4EMA 60', 'ATR0', 'ATR1', 'Variacion %1', 'ADX1', 'porcentaje a la media 11', 'diferencia0', 'cambio0'). Luego se usan con tres retrasos cada una.

### 4.4 Descripción de cada variable

Open: Precio de apertura de la sesión, en este trabajo la sesión hace referencia a una temporalidad diaria.

High: Precio más alto de la sesión.

Close: Precio del cierre de la sesión.

Low: Precio más bajo de la sesión.

Media de Bollinger: Media móvil simple de 19 períodos.

Banda Superior: La banda superior es la media de n días más el doble de la desviación media cuadrática de esa media (Bollinger & Commodities, 1992).

Banda Inferior: La banda inferior es la media de N días menos el doble de la desviación media al cuadrado de esa media, donde N se elige de forma que describa la tendencia a medio plazo (Bollinger & Commodities, 1992), en este caso N es igual a 19.

EMA: *Exponential Moving Average* (Media móvil exponencial): Las medias móviles exponenciales de la media móvil reducen el desfase aplicando más peso a los valores recientes. La ponderación aplicada al precio más reciente depende del número de períodos de la media móvil. El cálculo de una media móvil exponencial consta de tres pasos. En primer lugar, se calcula la media móvil simple. Una media móvil exponencial (EMA) tiene que empezar en algún lugar, por lo que se utiliza una media móvil simple como el período anterior (PRS) en el primer cálculo. En segundo lugar, se calcula el multiplicador de ponderación (MTL). En tercer lugar, se calcula la media móvil exponencial (Silva, Castilho, Pereira, & Brandao, 2014).

4EMA 3: Media móvil exponencial con ventana de 3 días.

4EMA 6: Media móvil exponencial con ventana de 6 días.

4EMA 13: Media móvil exponencial con ventana de 13 días.

4EMA 21: Media móvil exponencial con ventana de 21 días.

Máximo: Precio máximo por ventana de 45 días.

Mitad: Mitad del precio máximo por ventana de 45 días.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

Mínimos: Precio mínimo por ventana de 45 días.

Volumen: El volumen indica la cantidad de un instrumento financiero que se ha negociado.

Volumen MA: Volumen medio con ventana de 20 días.

ATR: Variación media entre los precios de cierre y apertura con una ventana de 3 días.

Variación %: Valor en porcentaje respecto al precio de apertura del valor absoluto de la diferencia entre el precio más alto y el más bajo del día.

Variación: Valor absoluto de la diferencia entre el precio más alto y el más bajo.

porcentaje a la media 1: Por ciento de la distancia del precio de cierre con respecto al media móvil de 30 períodos.

RSI: La fórmula RSI tiene en cuenta dos ecuaciones que intervienen en la resolución de la fórmula. La ecuación del primer componente obtiene el valor de fuerza relativa inicial (RS), que es la media de los cierres alcistas por encima de la media de los cierres bajistas, a lo largo de un período "N" representado en el siguiente cálculo: RS = Media móvil exponencial de 'N' períodos alcistas / Media móvil exponencial de 'N' períodos bajistas (en valor absoluto). El valor del indicador RSI se calcula indexando el indicador a 100 utilizando la siguiente fórmula: RSI = 100 - (100/1 + RS).

El RSI oscila entre valores de 0 a 100. Se considera que una acción está sobrecomprada cuando su RSI está por encima de 70, mientras que se considera sobrevendido cuando el RSI está por debajo de 30. Cuando el RSI está por encima de 50, indica una señal alcista, mientras que el valor se considera bajista cuando el RSI está por debajo de 50 (Chong & Ng, 2008).

Ángulo a la media: Calcula el porcentaje de cambio (tasa de cambio) entre el valor actual de SRC = (media móvil simple de 20 períodos) y su valor `length = 1` de barras históricas. Se calcula usando la siguiente fórmula: `100 * change (src, length) / src [length]`.

Stoch: El oscilador estocástico es un indicador de impulso que muestra la ubicación del cierre en relación con el rango alto-bajo durante un número determinado de períodos (Fidelity). Se manejan dos indicadores que son un oscilador estocástico rápido (%K) y un oscilador estocástico lento (%D), cuyas comparaciones son un buen indicador de la velocidad a la que los precios están cambiando o el impulso que tienen dichos precios. Sus fórmulas son:  $\%K=100 \times (C-Min) / Max-Min$   $\%D$  es la media móvil de %K de 3 períodos. Donde:

- C: valor del último cierre.
- Max: máximo del período de cálculo.
- Min: mínimo del período de cálculo y por defecto este período es de cinco.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

Si el precio de cierre se aproxima a los valores del mínimo para el período, %K disminuirá. Si dicho precio se aproxima al máximo, %K será cada vez más grande. Es así como el indicador informa sobre la situación del precio del cierre en relación con el máximo y mínimo del período dado para el cálculo.

ADX: El Índice de Movimiento Direccional Promedio (ADX) describe cuándo un mercado está en tendencia o no, es decir la fuerza de una tendencia. Cuando se combina con el *Plus Directional (+DI)* y el indicador direccional negativo (*-DI*), define la dirección de la tendencia. El objetivo principal del ADX es determinar si una acción está en tendencia o se encuentra en un rango (Silva et al., 2014).

Squeez: El *Squeeze Momentum Indicator* es una herramienta que funciona como oscilador de impulso; su función es indicar la explosividad y dirección con la que el precio del mercado se va a mover. La primera versión fue conocida como: “*TTM Squeeze*” de John Carter quien la explica en el capítulo 11 de su libro “*Mastering the Trade*” (Carter, 2012), posteriormente fue popularizada en la plataforma TradingView por un desarrollador llamado LazyBear quien recomienda usarla con un indicador ADX; al usar juntos estos dos indicadores se potencia la efectividad de los puntos de entrada y cierre de las posiciones.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

## Capítulo 5. Experimentación

En esta sección se abordan las métricas utilizadas. Además, se observa en detalle las configuraciones y las estructuras de los modelos. Se muestran los resultados obtenidos por cada activo estudiado, se analizan y discuten los resultados obtenidos.

### 5.1 Métricas de rendimiento para la evaluación de modelos de regresión

Para medir el desempeño de una regresión las métricas más usadas son:

- Error absoluto medio (MAE, *Mean Absolute Error*): Es el promedio de la diferencia absoluta entre el valor observado y los valores predichos. Se considera un puntaje lineal, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio. Por ejemplo, la diferencia entre 10 y 0 será el doble de la diferencia entre 5 y 0.
- Raíz del error cuadrático medio (RMSE, *Root Mean Squared Error*): Indica el ajuste absoluto del modelo a los datos, cuán cerca están los puntos de datos observados de los valores predichos del modelo y tiene la propiedad útil de estar en las mismas unidades que la variable de respuesta. Los valores más bajos de RMSE indican un mejor ajuste. Es una buena medida de la precisión con que el modelo predice la respuesta, y es el criterio más importante para ajustar si el propósito principal del modelo es la predicción (Gonzalez, 2018).
- Error porcentual absoluto medio (MAPE, *Mean Absolute Percentage Error*): Se utiliza a menudo en la práctica debido a su interpretación muy intuitiva en términos de error relativo. Su uso es relevante en finanzas, ya que las ganancias y las pérdidas suelen medirse en valores relativos. También es útil para calibrar los precios de los productos, ya que los clientes son a veces más sensibles a las variaciones relativas que a las absolutas (De Myttenaere, Golden, Le Grand, & Rossi, 2016).
- R cuadrado ( $R^2$ ): Representa la proporción de la varianza de ( $y$ ) que ha sido explicada por las variables independientes del modelo. Proporciona una indicación de la bondad del ajuste y, por tanto, una medida de la probabilidad de que las muestras no vistas sean predichas por el modelo, a través de la proporción de varianza explicada.

El MAE se calcula como se muestra en la ecuación 3. En la cual la variable  $y$  es el valor real,  $\hat{y}$  es el valor predicho,  $n$  es el número total de ejemplos.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i| \quad (3)$$

La RMSE se calcula como  $\sqrt{\text{MSE}}$  y el MSE se define como se muestra en la ecuación 4. En la cual la variable  $y$  es el valor real,  $\hat{y}$  es el valor predicho,  $n$  es el número total de ejemplos

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 \quad (4)$$

El MAPE se calcula como se muestra en la ecuación 5. En la cual la variable  $y$  es el valor real,  $\hat{y}$  es el valor predicho,  $n$  es el número total de ejemplos y  $\epsilon$  es un número arbitrario pequeño, pero estrictamente positivo para evitar resultados indefinidos cuando  $y$  es cero.

$$\text{MAPE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \frac{|y_i - \hat{y}_i|}{\max(\epsilon, |y_i|)} \quad (5)$$

$R^2$  se calcula como se muestra en la ecuación 6. En la cual la variable  $y$  es el valor real,  $\hat{y}$  es el valor predicho,  $n$  es el número total de ejemplos.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \text{ and } \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \epsilon_i^2. \quad (6)$$

## 5.2 Modelos de referencias

A continuación, se implementan dos modelos muy básicos de predicción con el objetivo no solo de evaluar los modelos usando métricas sino usando técnicas con un enfoque ingenuo y muy simple.

- Mañana igual que ayer: Esta es la técnica más simple que se usa, con un enfoque ingenuo, se predice que, el precio de cierre para el día ( $t$ ) será igual al del día ( $t-1$ ), pensado que no habrá una gran diferencia entre un día y el siguiente, este nos sirve de modelo base para comparar el rendimiento de las siguientes técnicas.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

- Ayer más variación: Técnica básica donde se predice que el precio de cierre para el día ( $t$ ) será igual a el precio de cierre del día ( $t-1$ ) más la diferencia entre los precios de cierre del día ( $t-1$ ) y ( $t-2$ ) pensando que, la tendencia se mantendrá de un día al siguiente y no debería cambiar significativamente.

### 5.3 Configuración de los modelos y resultados

En esta sección se explica la configuración de cada modelo y la arquitectura de cada una de las redes neuronales utilizadas y se explican en detalle los resultados obtenidos para la serie de precios del oro. En la próxima sección se muestran los mejores resultados obtenidos entre los 42 modelos para todas las series estudiadas.

Los modelos de aprendizaje profundo se implementaron utilizando la biblioteca *Keras*, mientras que los modelos de aprendizaje automático se implementaron utilizando la biblioteca *Scikit-learn*.

Las técnicas de predicción utilizadas en este trabajo son:

- Ayer más variación
- Básico, mañana igual que ayer
- Random forest
- Regresión lineal
- Regresión lineal 30-1
- Regresión lineal 200-1
- Red neuronal convolucional
- Red neuronal densa
- Red neuronal LSTM
- Red neuronal BLSTM

Partiendo de los diferentes conjuntos de variables mencionadas en el capítulo anterior, específicamente en la sección “Conjuntos de análisis”, se usan todas las técnicas de aprendizaje automático. Se tendrán ocho modelos de aprendizaje automático por cada uno de los 5 conjuntos de variables tomadas como regresores, en total serían 40 modelos más los dos básicos.

Se mostrarán en detalles todos los modelos para una base de datos de las cinco usadas, en este caso será el Oro. También se presenta, una tabla resumen de los rendimientos de todas las demás para no mostrar los 200 modelos, pues en configuración serían iguales por cada base de datos. En la Figura 12 vemos cómo se comporta el precio del Oro en temporalidad diaria desde 1980 hasta 2022. Aquí se puede apreciar que, en el conjunto de validación hay precios superiores a los del conjunto de entrenamiento. Y se verá como el desempeño de cada método ante tal situación.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

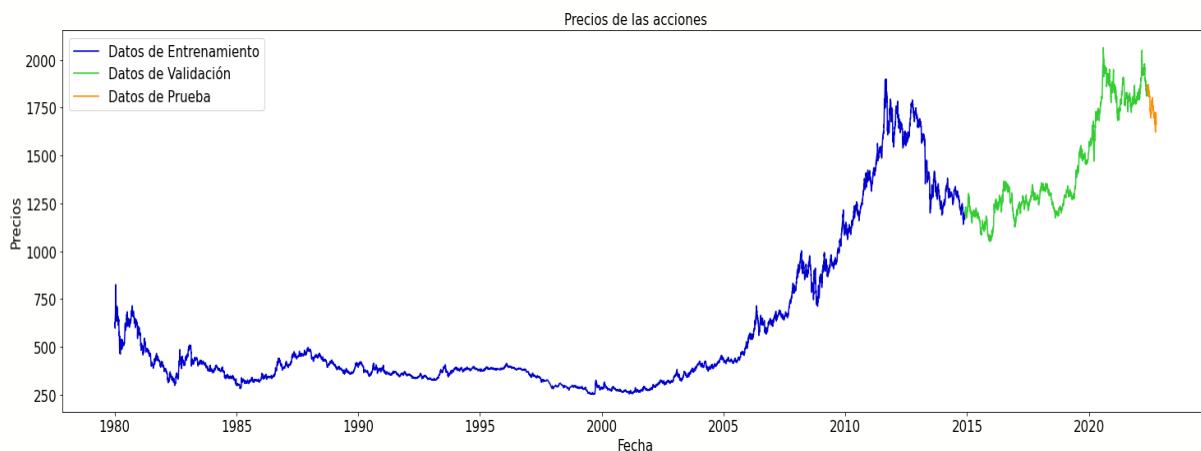


Figura 12. Precios de cierre en temporalidad diaria. Fuente: elaboración propia

### 5.3.1 Modelos

#### Ayer más variación

Este modelo es uno de los dos modelos básicos, en el cual se predice para el próximo día, utilizando la suma de la diferencia entre los dos días anteriores ( $t-1$ ) ( $t-2$ ) y el precio de cierre del día anterior ( $t-1$ ). Como este modelo no lleva un entrenamiento, se utilizará el mismo intervalo de tiempo para evaluar el modelo que los demás métodos, así se estarían comparando los resultados siempre sobre el mismo conjunto de datos.

En la Figura 13 se puede ver los precios reales representados con color rojo y los precios predichos en color azul. Por lo visto en la gráfica en la Figura 13, las dos series están superpuestas y representan una buena aproximación para el periodo de entrenamiento.

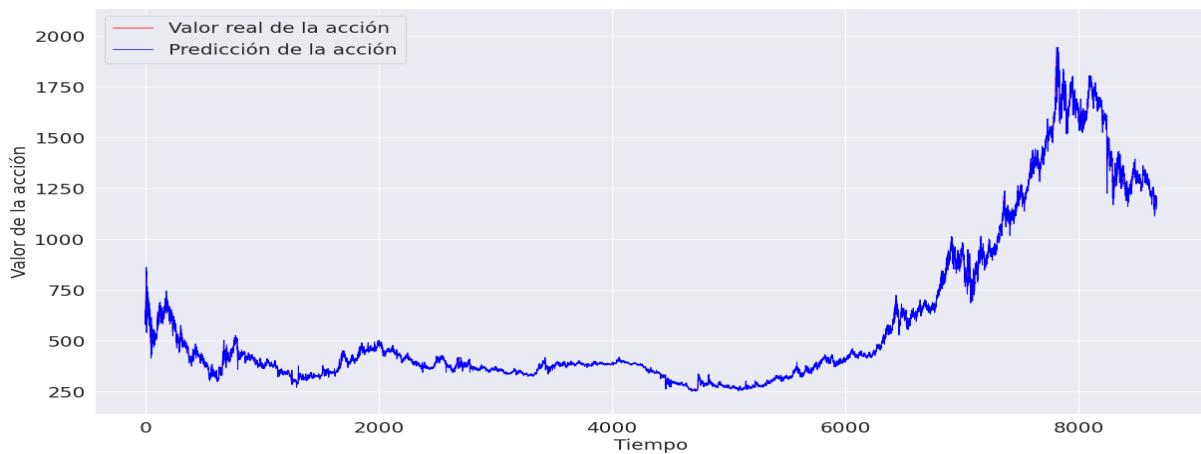


Figura 13. Precios reales y predichos en el periodo de entrenamiento. Fuente: elaboración propia.

En la Figura 14, se representa los precios reales con el color rojo y los precios predichos con el color azul para el periodo de validación.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

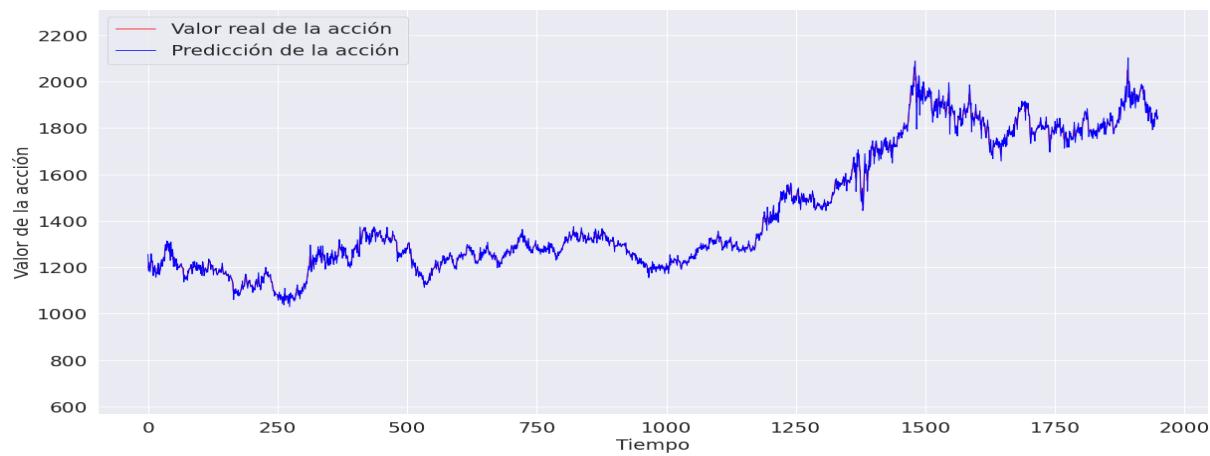


Figura 14. Precios reales y predichos en el periodo de validación. Fuente: elaboración propia.

Hasta este punto se puede notar que, aunque simple es una buena estrategia, como se observa en la Figura 14. Sin embargo, el comportamiento del algoritmo en un conjunto más pequeño como el de prueba, se puede ver que los valores predichos no están solapados (ver Figura 15).

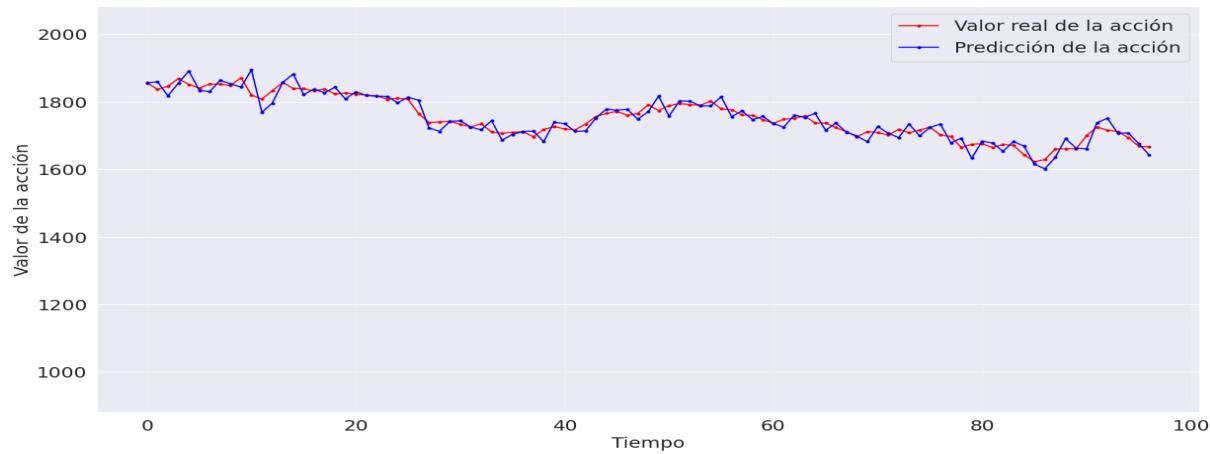


Figura 15. Precios reales y predichos en el periodo de prueba. Fuente: elaboración propia.

Una forma de analizar y evaluar los resultados es graficando el valor real con el predicho, en esta situación lo ideal sería esperar una función lineal con  $y=x$ , veremos qué tan cerca está el modelo básico de ese ideal a través de la Figura 16.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

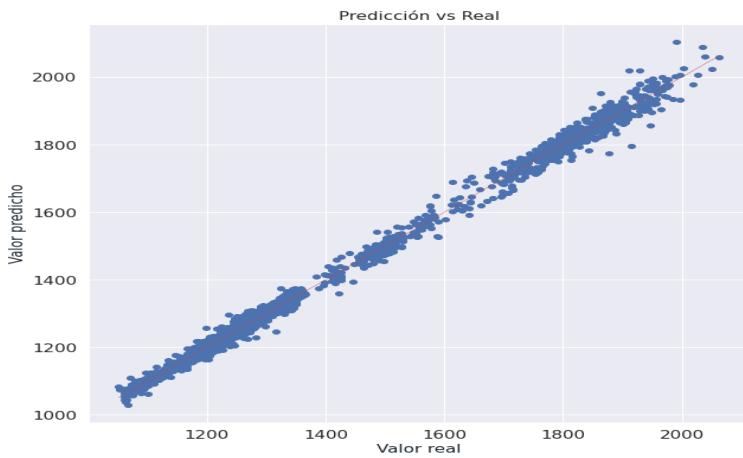


Figura 16. Precios reales vs predichos. Fuente: elaboración propia.

Al realizar una predicción se espera que el residuo (valor real – predicho) siga una distribución normal con media cero. Veremos el comportamiento para este método en la Figura 17.

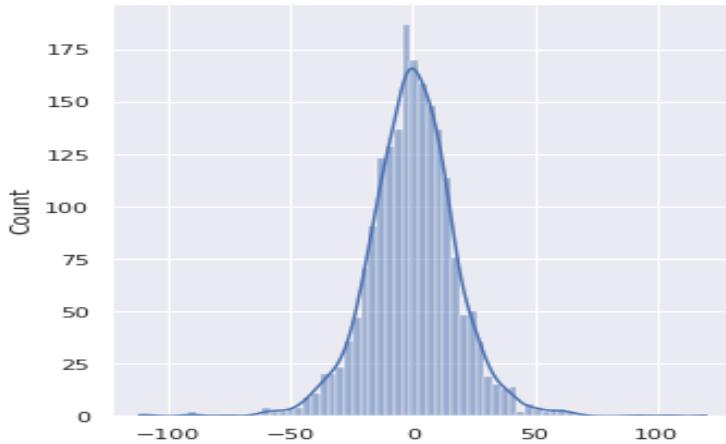


Figura 17. Distribución de los residuos. Fuente: elaboración propia.

Después de un análisis gráfico de algunos de los supuestos esperados para un modelo de regresión se puede decir que el primer método los cumple y es una buena aproximación de los valores reales, pero al realizar otros contrastes de hipótesis debemos rechazar el supuesto de normalidad de los errores.

Los estadísticos de asimetría (*Skewness*) y *Kurtosis* pueden emplearse para detectar desviaciones de la normalidad. Un valor de *Kurtosis* y/o coeficiente de asimetría entre -1 y 1, es generalmente considerada una ligera desviación de la normalidad (Bulmer, 1979). Entre -2 y 2 hay una evidente desviación de la normal pero no extrema.

Para este caso los valores del test son: *Kurtosis* 3.990 y *Skewness* -0.022 así que se rechaza la hipótesis de normalidad de los residuos, utilizando el método gráfico de cuantiles también se puede observar en la Figura 18 que no siguen una distribución normal.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

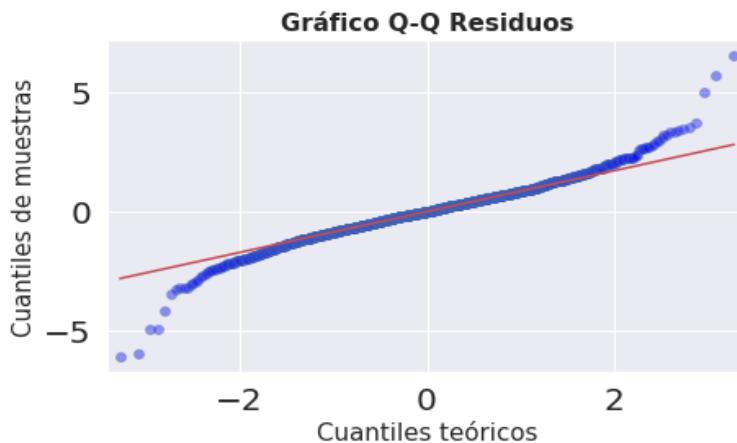


Figura 18. Cuantiles de los residuos. Fuente: elaboración propia.

Los test Shapiro-Wilk test y D'Agostino's K-squared test son dos de los test de hipótesis más empleados para analizar la normalidad. En ambos, se considera como hipótesis nula que los datos proceden de una distribución normal.

El p-value de estos test indica la probabilidad de obtener unos datos como los observados si realmente procediesen de una población con una distribución normal con la misma media y desviación que estos. Por lo tanto, si el p-value es menor que un determinado valor (típicamente 0.05), entonces se considera que hay evidencias suficientes para rechazar la normalidad.

El test de Shapiro-Wilk se desaconseja cuando se dispone de muchos datos (más de 50) por su elevada sensibilidad a pequeñas desviaciones de la normal (Rodrigo).

En la Tabla 6 se muestran los resultados de aplicar las pruebas de normalidad a los residuos.

Tabla 6. Normalidad de los residuos del modelo Ayer más variación. Fuente: elaboración propia.

	Estadístico	p_value	Normalidad
K-squared	184.463	8,80E-38	Falso
Shapiro-Wilk	0.965	2,35E-18	Falso

### Mañana igual que ayer

Este modelo simplemente predice que el precio de cierre del próximo día será igual al anterior. Utilizando la función `shift()` de la clase pandas se puede lograr fácilmente.

En la Figura 19 se observa la representación de los precios reales de color rojo; y los precios predichos de color azul. A simple vista parece una buena aproximación para el periodo de validación.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

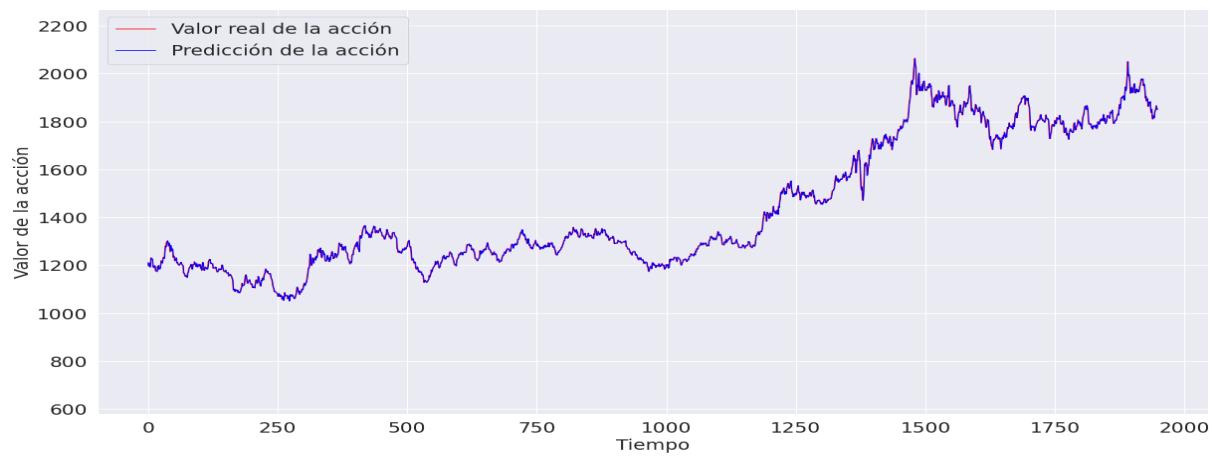


Figura 19. Precios reales y predichos en el periodo de validación. Fuente: elaboración propia.

Hasta este punto se pudiera interpretar como una buena estrategia, como se observa en la Figura 19. Sin embargo, el comportamiento del algoritmo en un conjunto más pequeño como el de prueba, se puede ver que los valores predichos no están solapados (ver Figura 20).

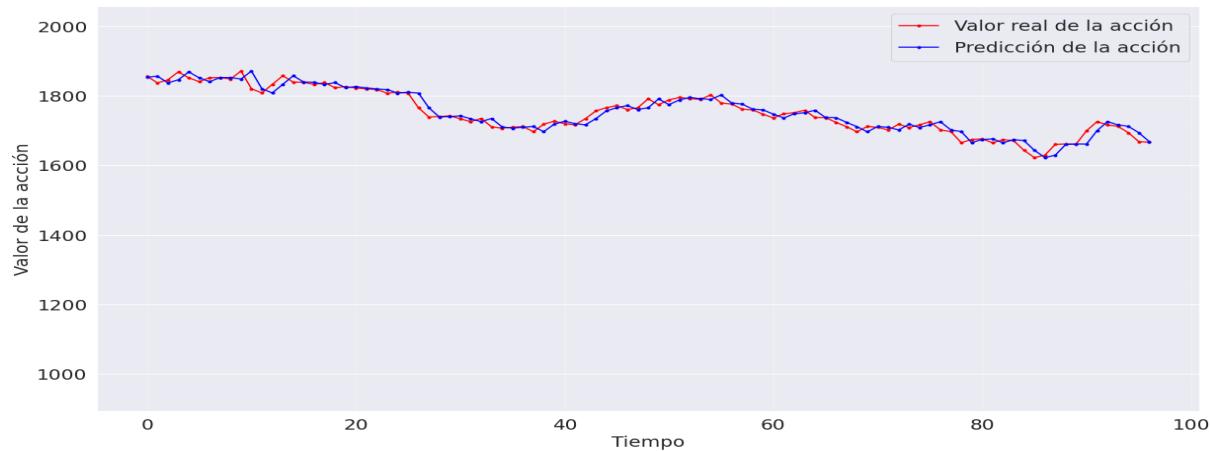


Figura 20. Precios reales y predichos en el periodo de prueba. Fuente: elaboración propia.

En la Figura 21 se compara el valor real con el predicho, en esta situación lo ideal sería ver una distribución de los puntos como una función lineal con  $y=x$ , se puede decir que el resultado es aceptable basándose en el análisis gráfico.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

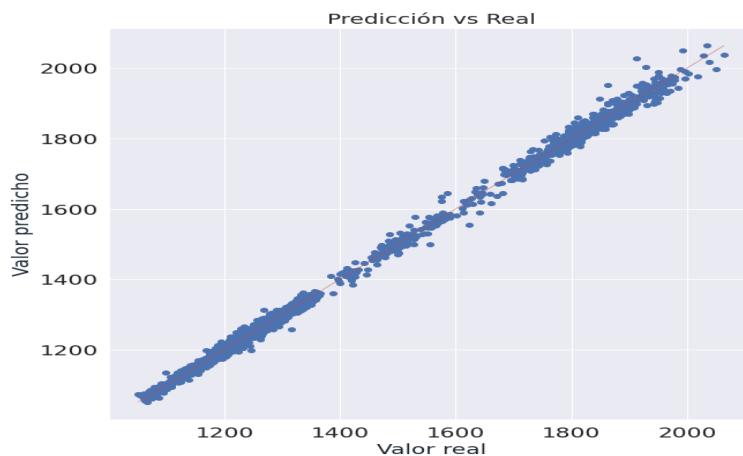


Figura 21. Precios reales vs predichos. Fuente: elaboración propia.

En la Figura 22 se muestra la distribución de los residuos, donde aparentemente sigue una distribución normal con media cero.

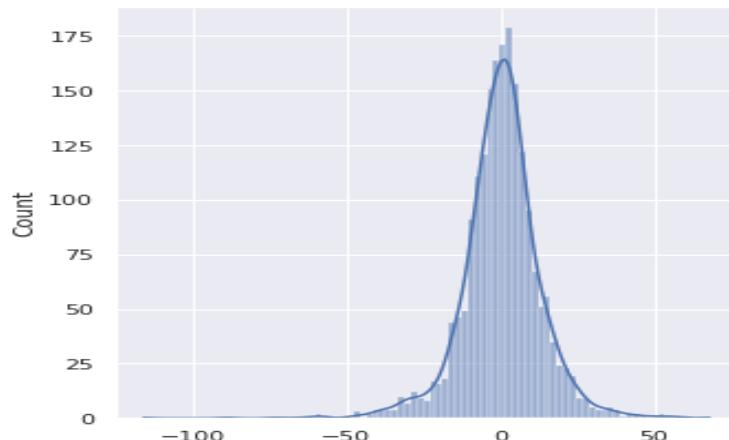


Figura 22. Distribución de los residuos. Fuente: elaboración propia.

Utilizando el método grafico de Cuantiles también se puede observar en la Figura 23 que no siguen una distribución normal.

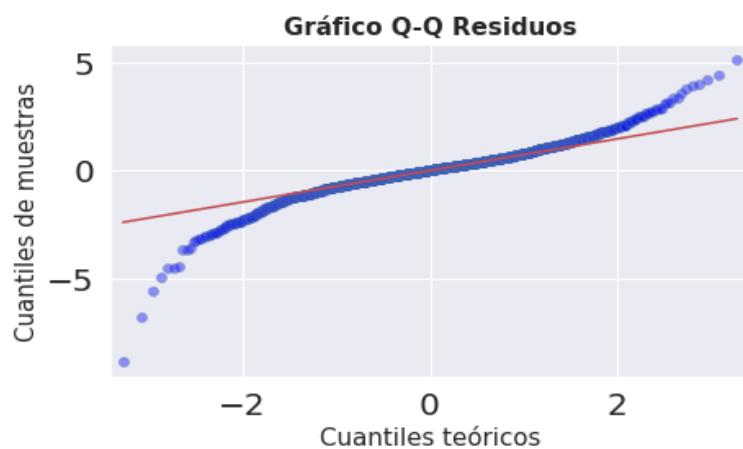


Figura 23. Cuantiles de los residuos. Fuente: elaboración propia.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

En la Tabla 7 se muestran los resultados de aplicar las pruebas de normalidad a los residuos.

Tabla 7. Normalidad de los residuos del modelo Mañana igual que ayer. Fuente: elaboración propia.

	Estadístico	p_value	Normalidad
K-squared	406.676	4.914e-89	Falso
Shapiro-Wilk	0.933	9.574e-29	Falso

### 5.3.2 Modelos de aprendizaje automático utilizando el primer conjunto de datos Modelo Random Forest

Para la implantación de este modelo se utiliza la biblioteca sklearn, que tiene varias clases que facilitan el uso de algoritmos de aprendizaje automático, entre ellas RandomForestRegressor. Un bosque aleatorio es un meta-estimador que ajusta una serie de árboles de decisión clasificatorios a varias sub-muestras del conjunto de datos y utiliza el promedio para mejorar la precisión predictiva y controlar el sobreajuste. El tamaño de la sub-muestra se controla con el parámetro `max_samples` si `bootstrap=True` (predeterminado); de lo contrario, se usa todo el conjunto de datos para construir cada árbol.

En este caso se usan los atributos por defectos de la clase y solo se establece la cantidad de características igual a la cantidad de variables independientes, multiplicado por el número de retrasos. El modelo se entrena utilizando el método `fit()` de la clase que recibe como parámetros la matriz de variables independientes con sus respectivos retrasos por tantas filas como días tengo de entrenamiento y el vector con los precios de cierres. La forma de la matriz de variables independientes es de dimensión 8666 por 48 y la del vector con los precios de cierres es 8666. Utilizando el método `predict()`, se predicen los valores para los datos de validación, pasando como parámetro la matriz con los datos de validación con dimensión 1949 por 48 y devuelve un array de longitud 1949 con los precios predichos. Luego utilizando los valores reales del close del periodo de validación y los valores predichos se calculan las métricas: *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), *Mean Absolute Percentage Error* (MAPE) y  $R^2$ .

En la Figura 24 se muestra la representación de los precios reales, que es la línea roja; y los precios predichos, que es la línea azul. Por lo visto parece una buena aproximación para el periodo de entrenamiento.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

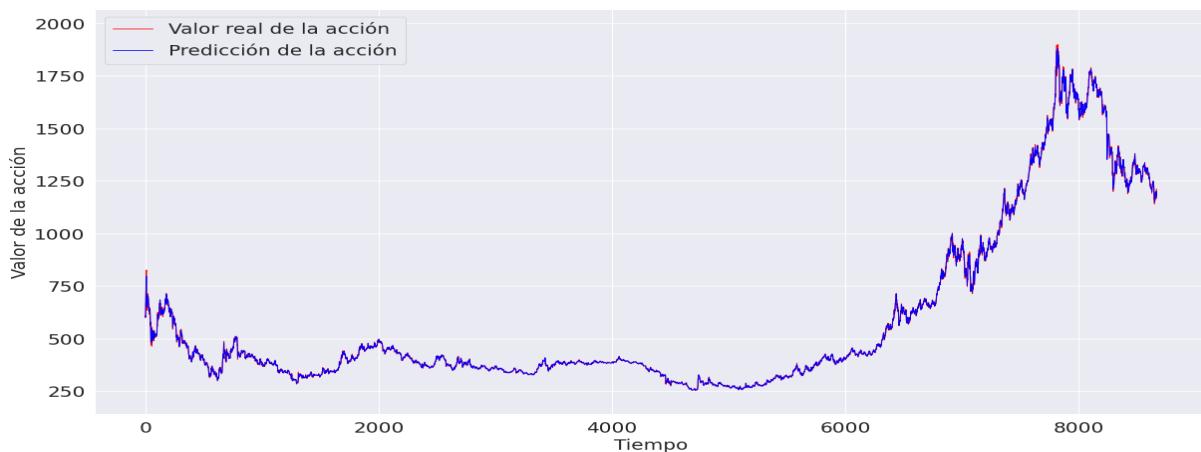


Figura 24. Precios reales y predichos en el periodo de entrenamiento. Fuente: elaboración propia

Para el periodo de validación se representan en la Figura 25 los precios reales con el color rojo y los precios predichos con el color azul. Se observa que el modelo no predice relativamente bien los precios donde los precios de cierre son superiores a los del periodo de validación.

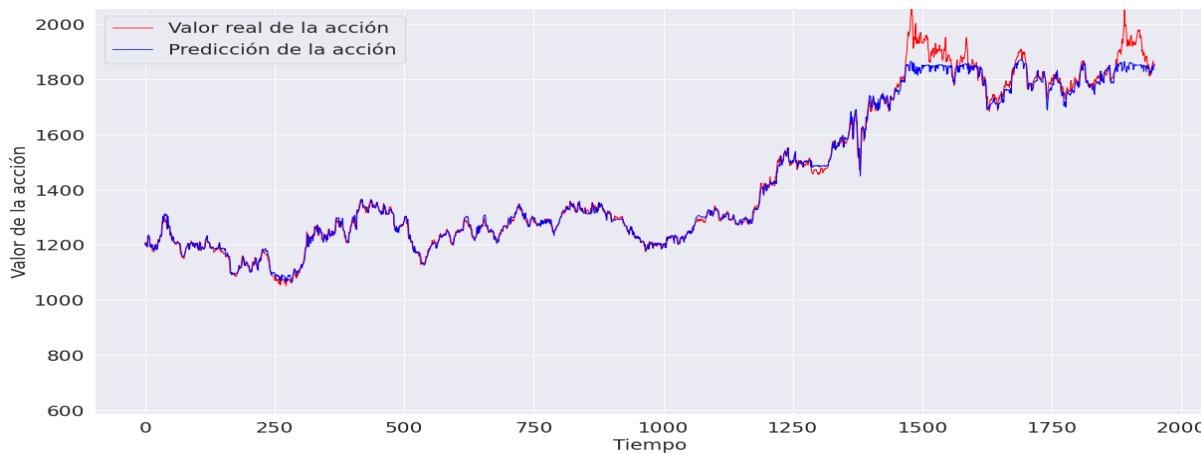


Figura 25. Precios reales y predichos en el periodo de validación. Fuente: elaboración propia.

Mostrando una gráfica con el precio real contra el precio predicho se espera que se comporte como una función lineal  $y=x$ . En la Figura 26 se observa que el modelo no predice valores superiores a los vistos en la etapa de entrenamiento. Cuando los precios reales son superiores a 1900 el modelo solo predice valores cercanos a 1800, siendo este uno de los problemas de los modelos basados en árboles de decisión, donde se produce un sobreajuste y no se logra generalizar para datos fuera del conjunto de entrenamiento.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

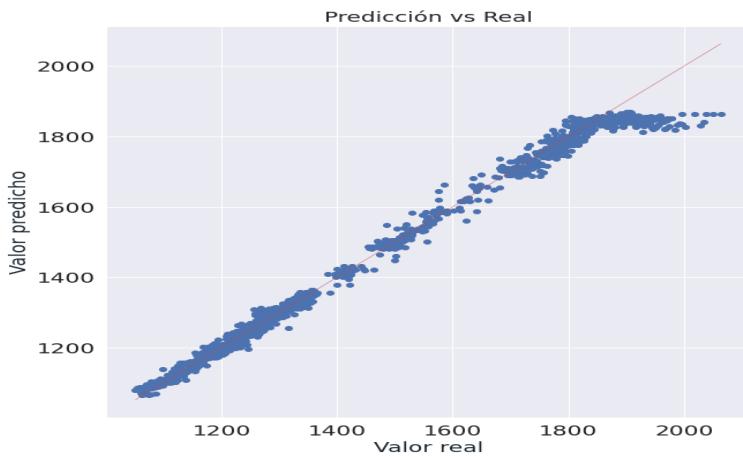


Figura 26. Precios reales contra precios predichos. Fuente: elaboración propia.

Otra forma visual de valorar el comportamiento de los valores predichos es representando los residuos (valores reales – predichos) en un histograma a gráfica de distribución. Estos residuos no siguen una distribución normal. En la Figura 27 se muestra en un histograma de los residuos.

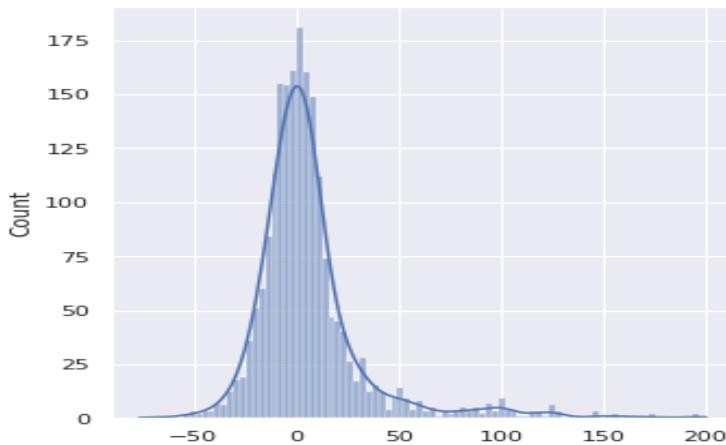


Figura 27. Distribución de los residuos. Fuente: elaboración propia

## Modelo de Regresión lineal

Para la implantación de este modelo se utiliza la biblioteca `sklearn` que tiene varias clases que facilitan el uso de algoritmos de aprendizaje automático, entre ellas `LinearRegression`. El modelo se entrena utilizando el método `fit()` de la clase que recibe como parámetros la matriz de variables independientes, con sus respectivos retrasos por tantas filas como días tengo de entrenamiento y el vector con los precios de cierres.

La forma de la matriz de variables independientes es de dimensión 8666 por 48 y la del vector con los precios de cierre es de 8666. Utilizando el método `predict()`, se predicen los valores para los datos de validación, se pasa como parámetro la matriz con los datos de validación con la forma 1949 por 48 y devuelve un array con longitud 1949 con los precios predichos. Luego utilizando los valores reales del `close` del periodo de validación y los valores predichos, se calculan las métricas: *Mean Absolute*

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

*Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) y R<sup>2</sup>.*

En la Figura 28 se muestran los precios reales, en color rojo; y los precios predichos en azul. A simple vista parece una aproximación muy buena para el periodo de entrenamiento.

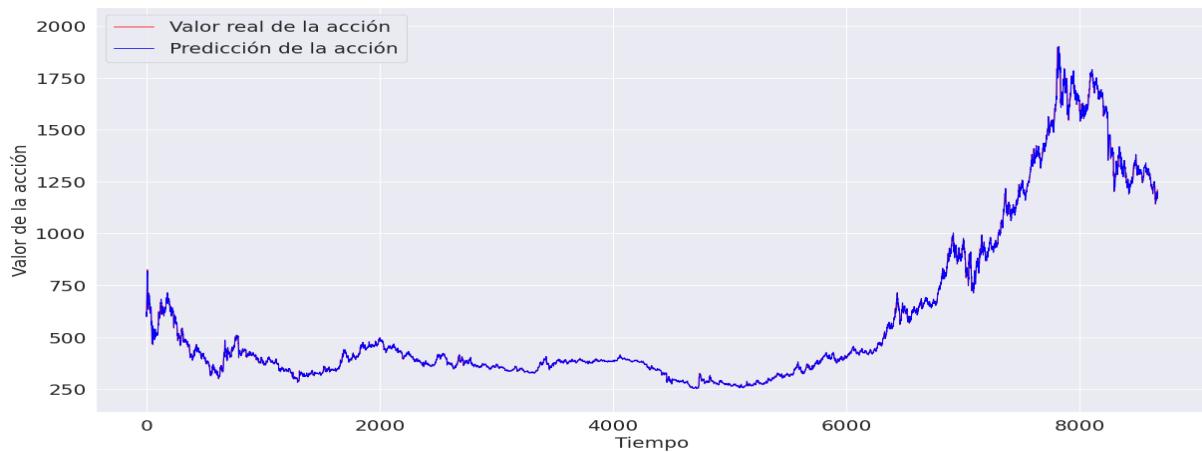


Figura 28. Precios reales y predichos en el periodo de entrenamiento. Fuente: elaboración propia.

En la Figura 29 se representan los precios reales con una línea roja y los precios predichos con una línea azul. Parece una buena aproximación para el periodo de validación.

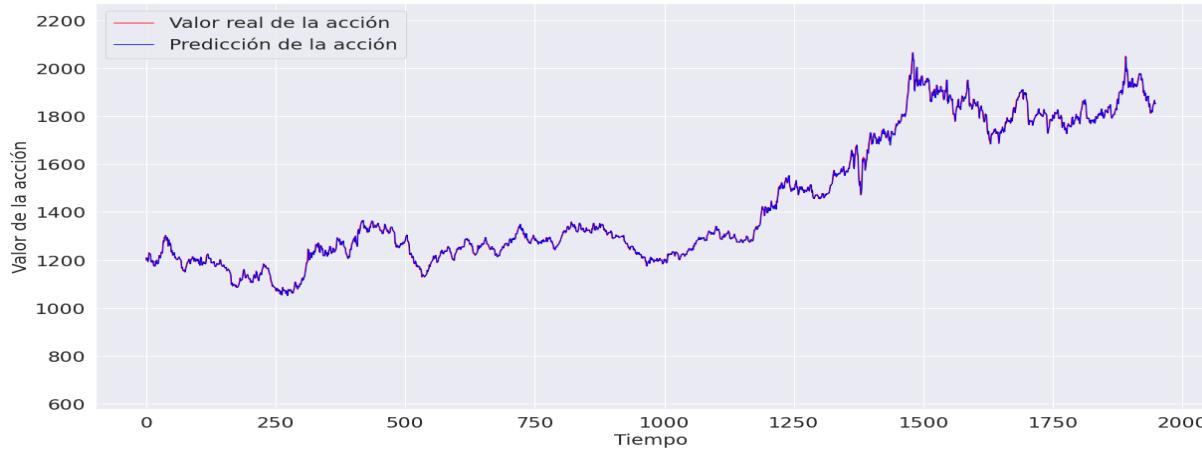


Figura 29. Precios reales y predichos en el periodo de validación. Fuente: elaboración propia.

En la Figura 30 se muestra el valor real contra el precio predicho, en esta situación lo ideal sería ver una distribución de los puntos como una función lineal con  $y=x$ .

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

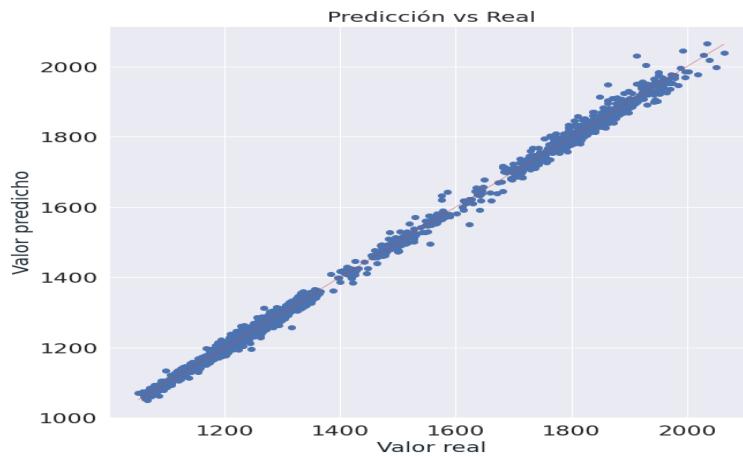


Figura 30. Precios reales contra los precios predichos. Fuente: elaboración propia.

En la Figura 31 se muestra la distribución de los residuos. Se puede notar que se asemejan a una distribución normal con media cero. Sin embargo analizando los residuos por el método gráfico de cuantiles (ver Figura 32) podemos rechazar la hipótesis de normalidad de los residuos y así también los confirman las pruebas de normalidad que se muestran en la Tabla 8.

Tabla 8. Normalidad de los residuos. Fuente: elaboración propia.

	Estadístico	p_value	Normalidad
K-squared	412.852	2.240e-90	Falso
Shapiro-Wilk	0.933	1.117e-28	Falso

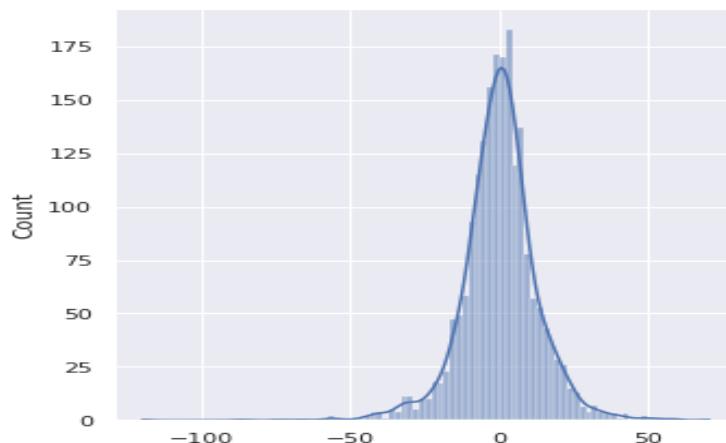


Figura 31. Distribución de los residuos. Fuente: elaboración propia.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

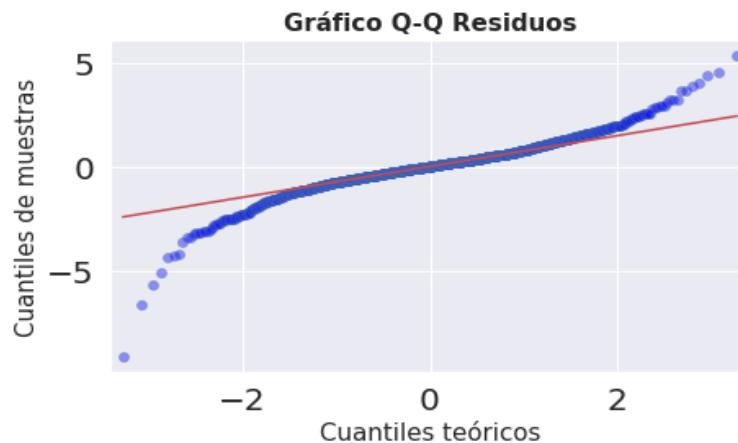


Figura 32. Cuantiles de los residuos. Fuente: elaboración propia.

## Modelos de redes neuronales

### Modelo de red denso

Este modelo se implementa usando la librería de keras, la red utiliza capas de tipo densas con función de activación Relu y como optimizador se usa Adam. La estructura de la red se muestra en la Figura 33. En este modelo se usan múltiples combinaciones posibles de los parámetros, como por ejemplo la probabilidad de *Dropout* en las capas intermedias, tasa de aprendizaje (*learning rate*) y tamaño del Batch. Así se prueban un total de 18 combinaciones y se selecciona el mejor modelo basado en el menor error cuadrático medio.

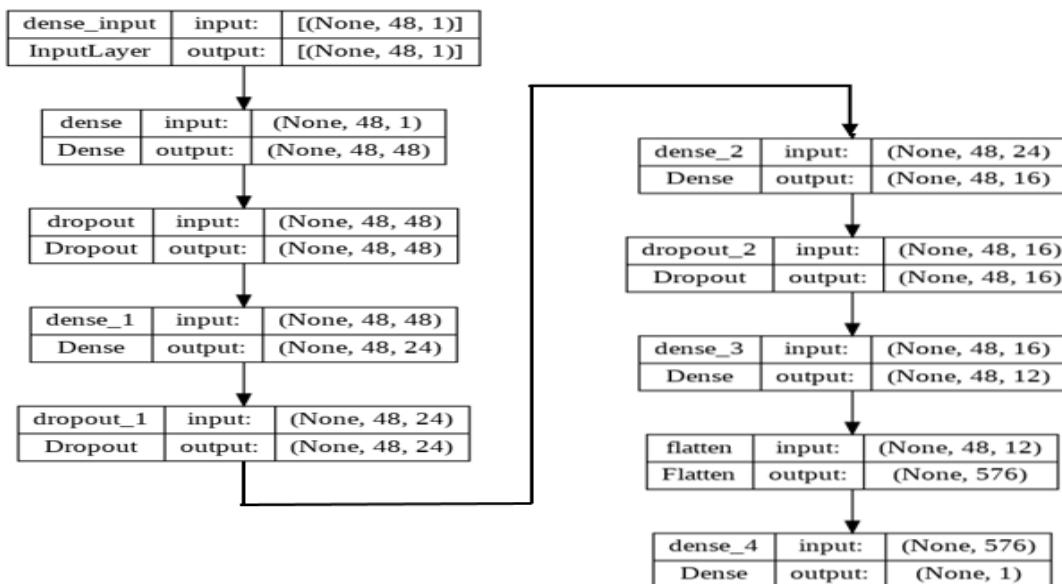


Figura 33. Estructura de la densa. Fuente: elaboración propia.

En la Figura 34 se muestra la predicción para el periodo de validación. La línea roja representa el precio real y la azul el precio predicho.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

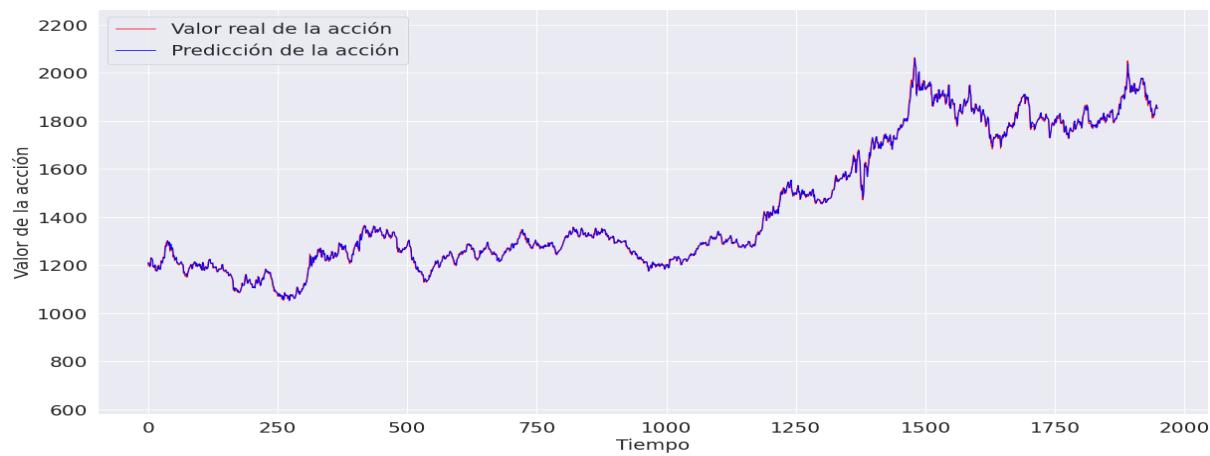


Figura 34. Precios reales y predichos en el periodo de validación. Fuente: elaboración propia.

En la Figura 35 se muestra el valor real en comparación con el predicho, en esta situación lo ideal sería ver una distribución de los puntos como una función lineal con  $y=x$ , se puede decir que el resultado es bastante bueno basado en este análisis gráfico.

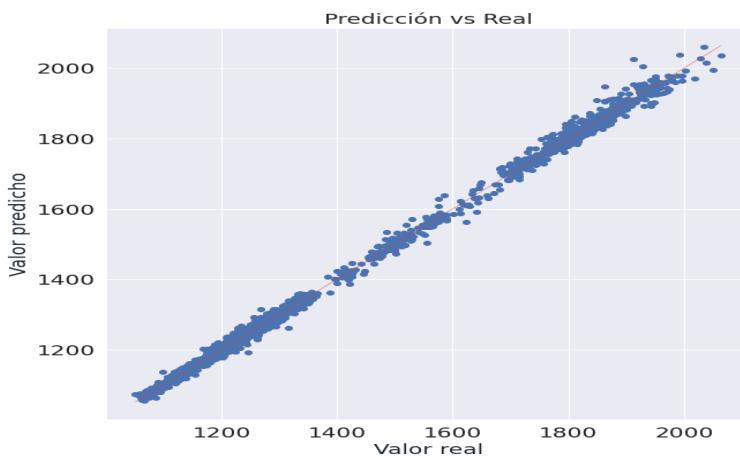


Figura 35. Precios reales contra el precio predicho. Fuente: elaboración propia.

En la Figura 36 se muestra la distribución de los residuos y al igual que los modelos anteriores la distribución de los errores se asemeja a una distribución normal pero en ninguno de los casos lo fue.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

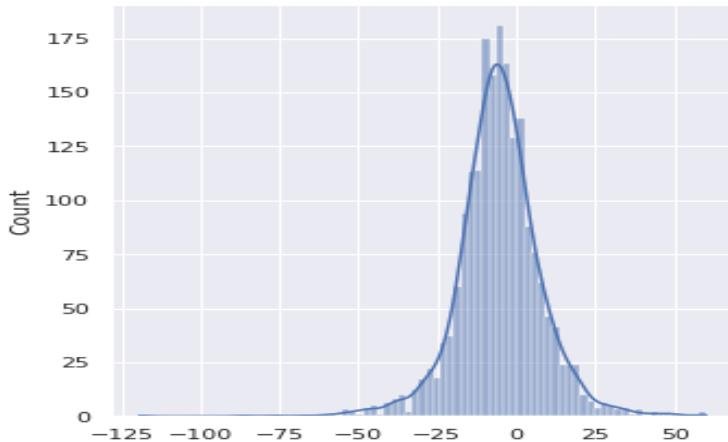


Figura 36. Distribución de los residuos. Fuente: elaboración propia.

En la Tabla 9 se muestran los resultados de las pruebas de normalidad aplicadas a los residuos.

Tabla 9. Normalidad de los residuos. Fuente: elaboración propia.

	Estadístico	p_value	Normalidad
K-squared	330.080	2.109e-72	Falso
Shapiro-Wilk	0.945	1.872e-26	Falso

## Modelo de red Convolucional

En este modelo se usan múltiples combinaciones posibles de los parámetros, *dropout* en las capas intermedias, tasa de aprendizaje (*learning rate*) y tamaño del *batch*. Como se muestra a continuación: num\_nodes [16, 32, 64], dropout\_prob [0, 0.2], lr [0.01, 0.005, 0.001], batch\_size in [32, 64, 128]. En total se probaron 54 combinaciones posibles de este modelo para seleccionar el mejor. La métrica para elegir el mejor modelo entre los 54 modelos, fue el error cuadrático medio.

La arquitectura de la red se muestra en la Figura 37.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

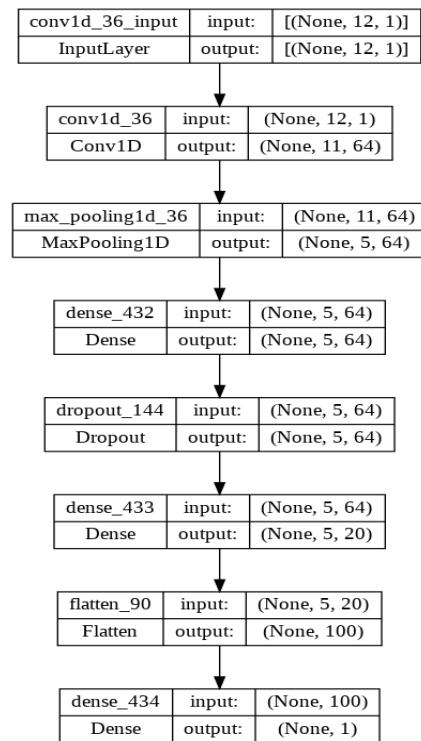


Figura 37. Arquitectura de red Convolucional. Fuente: elaboración propia.

En la Figura 38 se muestra la predicción para el periodo de validación. La línea roja representa el precio real y la azul el precio predicho.



Figura 38. Precios reales y predichos en el periodo de validación. Fuente: elaboración propia.

En la Figura 39 se muestra el valor real en comparación con el predicho, en esta situación lo ideal sería ver una distribución de los puntos como una función lineal con  $y=x$ , se puede decir que el resultado es bueno basado en este análisis gráfico.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

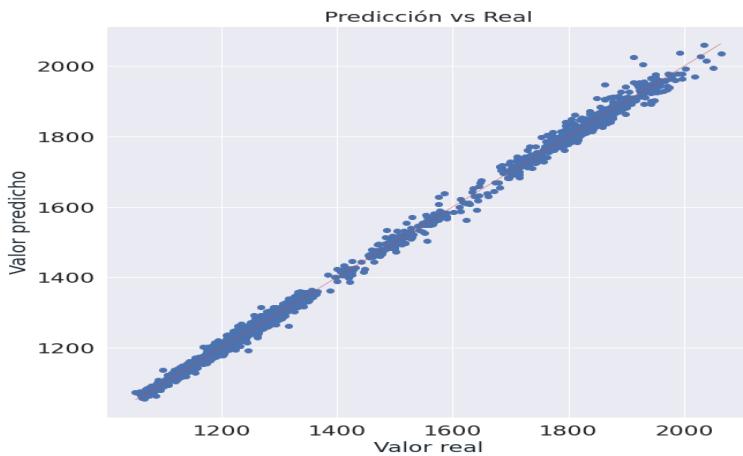


Figura 39. Precios reales contra precios predichos. Fuente: elaboración propia.

En la Figura 40 se muestra la distribución de los residuos. Se asemeja a una distribución normal con media cero, pero después de realizar las pruebas de normalidad de los residuos no se puede aceptar la hipótesis de normalidad de los residuos.

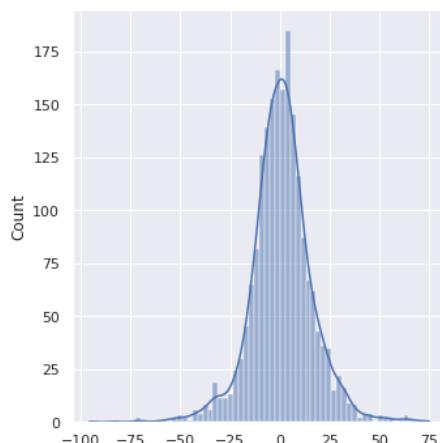


Figura 40 . Distribución de los residuos. Fuente: elaboración propia.

En la Tabla 10 se muestran los resultados de las pruebas de normalidad de los residuos para el modelo de red convolucional.

Tabla 10. Normalidad de los residuos. Fuente: elaboración propia.

	Estadístico	p_value	Normalidad
K-squared	162.985	4.058e-36	Falso
Shapiro-Wilk	0.961	1.664e-22	Falso

## Modelo red recurrente LSTM

En este modelo se usan múltiples combinaciones posibles de los parámetros, cantidad de neuronas en la capa de entrada y tamaño del batch. Como se muestra a continuación: `num_node [32, 64, 128]`, `batch_size [32, 64, 128]`. En total se probaron 6 combinaciones posibles de este modelo para seleccionar el mejor

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

con 400 épocas de entrenamiento. La variable para elegir el mejor modelo entre los seis posibles modelos, fue el error cuadrático medio.

Estructura de la red LSTM se muestra en la Figura 41.

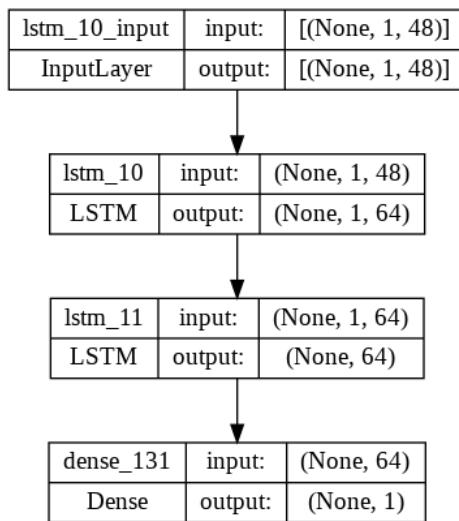


Figura 41. Estructura de la red LSTM. Fuente: elaboración propia

En la Figura 42 se muestra la predicción para el periodo de validación. De color rojo el precio real y en azul el precio predicho.



Figura 42 . Precios reales y predichos en el periodo de validación por el modelo LSTM. Fuente: elaboración propia.

La Figura 43 muestra el valor real en comparación con el predicho, en esta situación lo ideal sería ver una distribución de los puntos como una función lineal con  $y=x$ , se puede decir que el resultado es bueno según el análisis gráfico.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

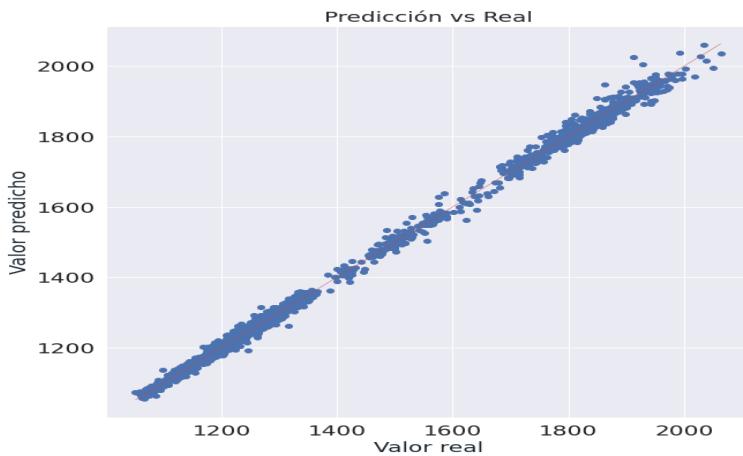


Figura 43 . Precios reales contra los precios predichos del modelo LSTM. Fuente: elaboración propia.

En la Figura 44 se muestra la distribución de los residuos del modelo LSTM. Se puede notar que sigue una distribución normal con media cero como se espera de una buena predicción, en contraposición, otra vez las pruebas de normalidad demuestran que no se puede aceptar el supuesto de normalidad de los residuos como se puede ver en los resultados que muestra la Tabla 11.

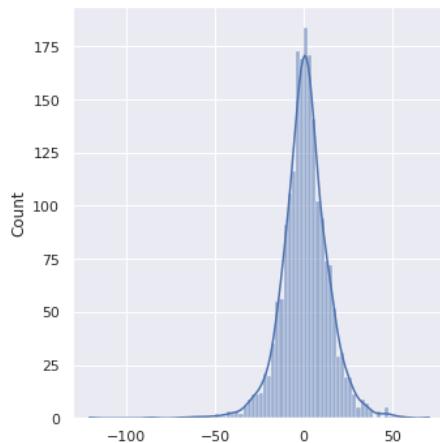


Figura 44 . Distribución de los residuos modelo LSTM. Fuente: elaboración propia.

Tabla 11.. Normalidad de los residuos. Fuente: elaboración propia.

	Estadístico	p_value	Normalidad
K-squared	330.080	2.109e-72	Falso
Shapiro-Wilk	0.945	1.872e-26	Falso

### Modelo red recurrente LSTM bidireccional (BLSTM)

En este modelo se usan múltiples combinaciones posibles de los parámetros, número de neuronas en la primera capa, tasa de aprendizaje (*learning rate*) y tamaño del *batch*. Como se muestra a continuación: `num_nodes [50, 75, 100]`, `lr [0.01, 0.005, 0.001]`, `batch_size in [32, 64, 128]`. En total se probaron 27 combinaciones posibles de este modelo para seleccionar el mejor. La variable para

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

elegir el mejor modelo entre los 27 posibles fue el error cuadrático medio. En la Figura 45 se observa la estructura del modelo utilizado.

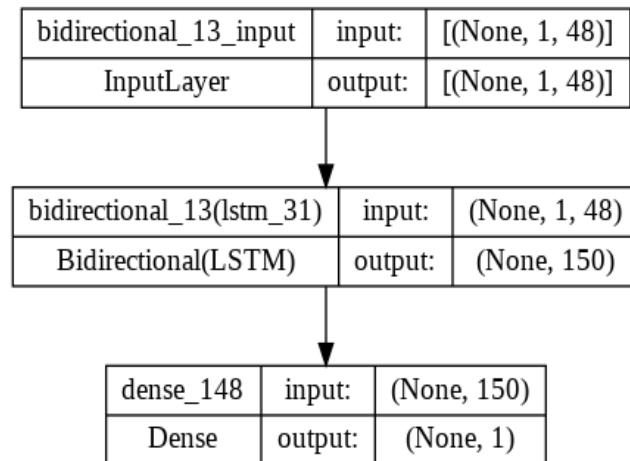


Figura 45. Estructura de la red BLSTM. Fuente: elaboración propia.

En la Figura 46 se muestra la predicción para el periodo de validación. La línea roja representa el precio real y la azul representa el precio predicho.

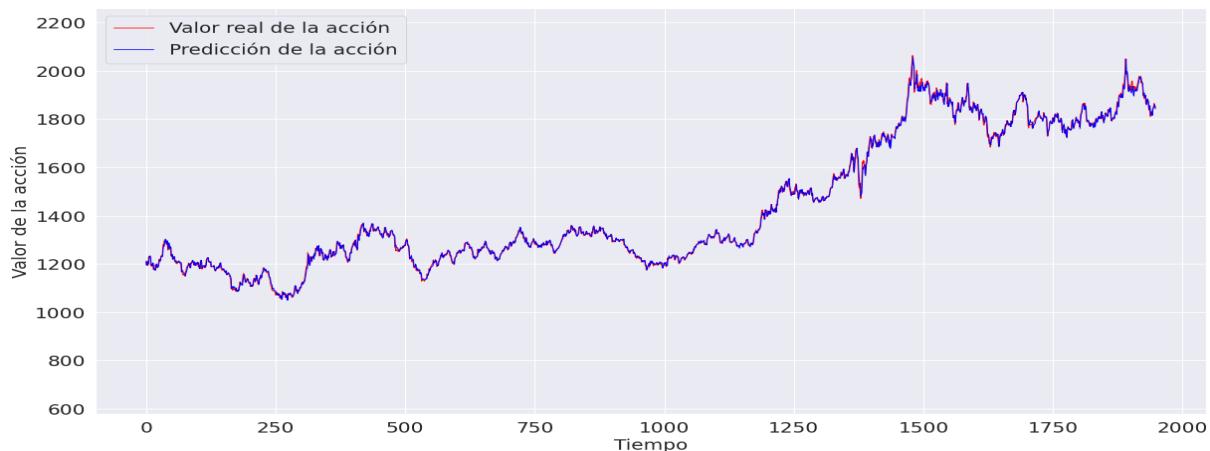


Figura 46. Precios reales y predichos en el período de validación. Fuente: elaboración propia.

En la Figura 47 se muestra el valor real comparado con valor el predicho, lo ideal sería ver una distribución de los puntos como una función lineal con  $y=x$ , se puede decir que el resultado es una buena aproximación basado en este análisis gráfico.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

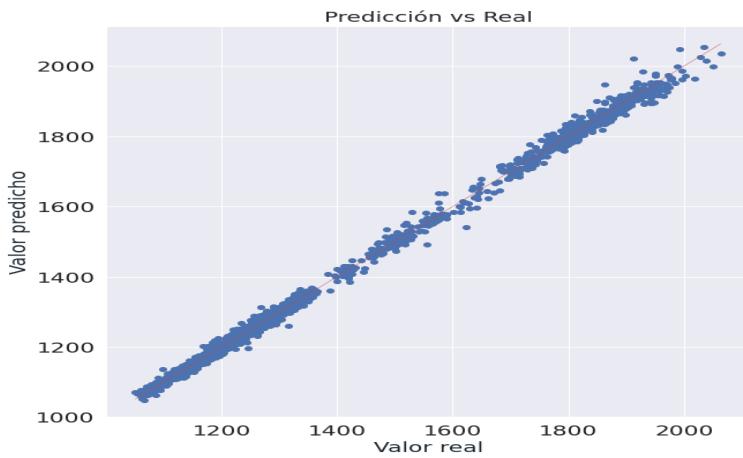


Figura 47. Precios reales vs predichos por el modelo BLSTM. Fuente: elaboración propia.

En la Figura 48 se muestra la distribución de los residuos. Parece seguir una distribución normal con media cero como los modelos anteriores, esto es lo deseado pues los errores deberían tener varianza constante y estar cercanos a cero, sin embargo, las pruebas de normalidad rechazan la hipótesis de normalidad de los residuos como se puede observar en los resultados mostrados en la Tabla 12.

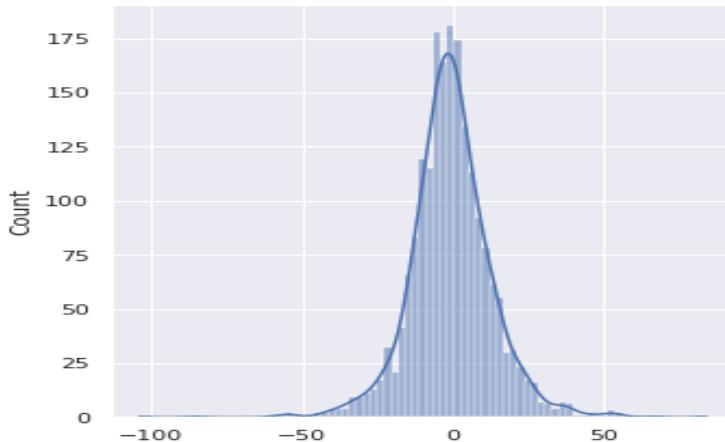


Figura 48. Distribución de los residuos del modelo BLSTM. Fuente: elaboración propia.

Tabla 12. Normalidad de los residuos del modelo BLSTM. Fuente: elaboración propia

	Estadístico	p_value	Normalidad
K-squared	234.414	1.252e-51	Falso
Shapiro-Wilk	0.952	1.017e-24	Falso

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

Comprobando el supuesto de normalidad de los residuos de forma gráfica con el método de cuantiles ver Figura 49 se pudiera afirmar que no siguen una distribución normal.

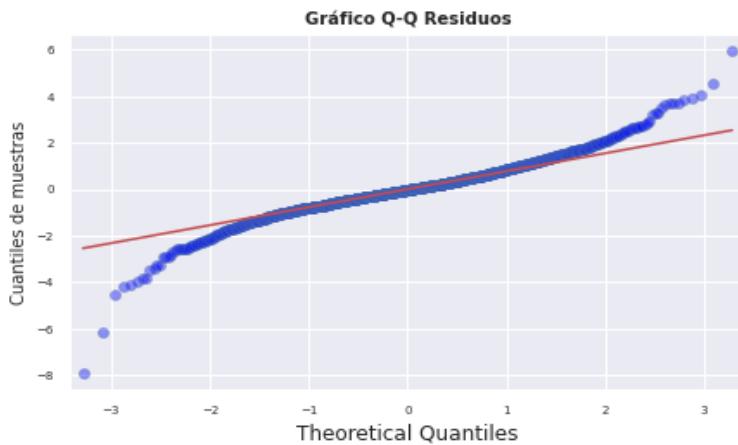


Figura 49 . Cuantiles de los residuos. Fuente: elaboración propia.

En otro análisis gráfico de los residuos podemos ver que los residuos no tienen varianza constante y se alejan de una media igual a cero como se puede ver en la Figura 50.

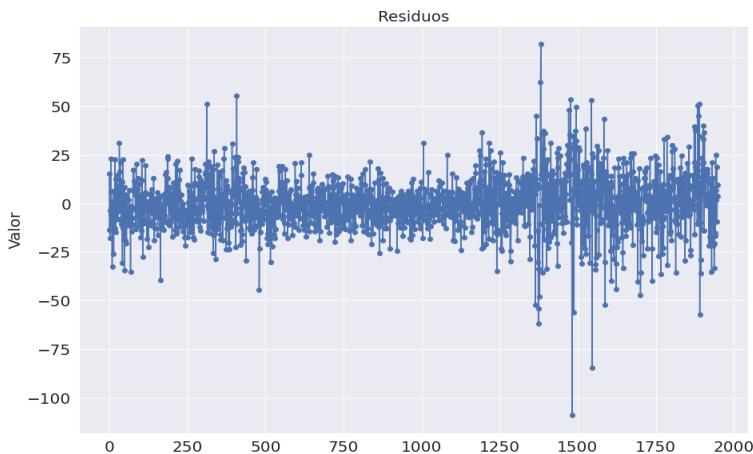


Figura 50. Residuos del modelo BLSTM. Fuente: Elaboración propia.

Al tratar con series temporales, decidir con qué datos entrenar y cuales usar para validación se presentan algunos problemas. Por ejemplo, en este caso se usa un 80% para entrenar, estos datos son la etapa inicial de la serie y se valida con un 19% de los datos, que están desplazados casi cinco años después de nuestro conjunto de entrenamiento. Por tanto, las predicciones se estarían haciendo con valores muy distantes en el tiempo.

Para intentar solucionar de alguna forma esta problemática, se decide realizar dos experimentos más. Utilizando las variables de entrada de cada uno de los 5 conjuntos se crean un primer modelo de regresión lineal que entrena con 30 días y valida con el día 31. Así se corre esa ventana de datos de 30 días hasta barrer todo el periodo

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

de entrenamiento inicial que son 8666 datos, luego para calcular las métricas de evaluación se utiliza el mismo periodo que el de todos los modelos anteriores para ser consistente y estricto en la comparación de los resultados. El segundo modelo sigue la misma filosofía que el anterior con ventana de 200 días.

Hasta ahora se ha visto como los modelos hacen una predicción bastante cercana al precio real al menos gráficamente. En la Figura 51 se muestra una comparación por diferentes métricas de todos los modelos para el primer conjunto de variables estudiado.

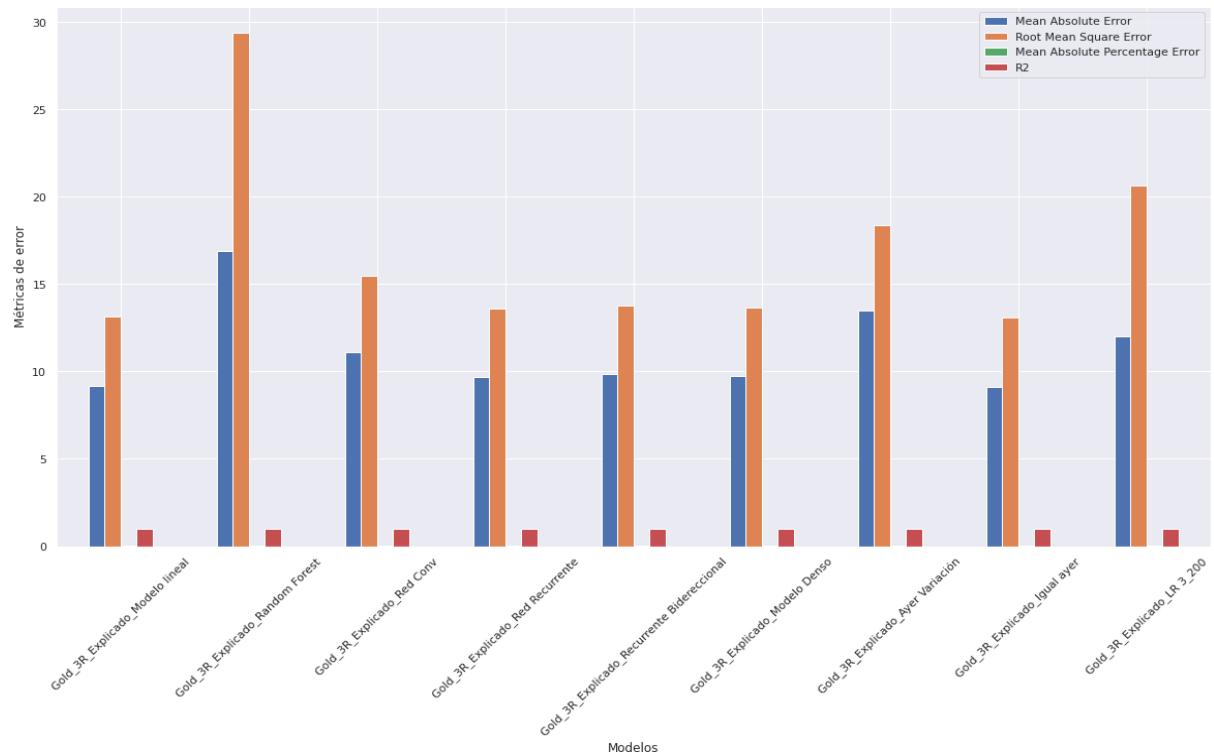


Figura 51. Métricas de error de los modelos utilizando el conjunto de datos uno. Fuente: elaboración propia.

### 5.3.3 Resultados obtenidos

En la Tabla 13 se muestran los modelos ordenados de forma descendente por el error medio absoluto utilizando el primer conjunto de datos. En la que se observa que el mejor modelo es el modelo simple, Mañana igual que ayer.

Tabla 13. Modelos ordenados utilizados en el primer conjunto de datos. Fuente: elaboración propia.

Modelo	MAE	RMSE	MAPE	R2
<b>Mañana igual que ayer</b>	9.128	13.119	0.006	0.998
<b>Regresión lineal</b>	9.150	13.148	0.006	0.998
<b>Red neuronal densa</b>	9.397	13.318	0.006	0.998
<b>Red neuronal LSTM</b>	9.683	13.602	0.007	0.997
<b>Red neuronal BLSTM</b>	9.773	13.708	0.007	0.997
<b>Red Convolucional</b>	11.129	15.465	0.008	0.997
<b>Regresión lineal 200-1</b>	12.020	20.625	0.008	0.994

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>Ayer más variación</b>	13.483	18.354	0.009	0.995
<b>Random Forest</b>	16.878	29.408	0.011	0.988
<b>Regresión lineal 30-1</b>	48.716	85.591	0.033	0.900

En la Tabla 14 se muestran los modelos ordenados descendenteamente por el error medio absoluto utilizando el segundo conjunto de datos. Obteniendo como mejor modelo la Regresión lineal.

Tabla 14. Modelos ordenados utilizados en el segundo conjunto de datos. Fuente: elaboración propia.

Modelo	MAE	RMSE	MAPE	R2
<b>Regresión lineal</b>	9.141	13.112	0.006	0.998
<b>Red neuronal densa</b>	9.599	13.573	0.007	0.997
<b>Red neuronal BLSTM</b>	9.688	13.614	0.007	0.997
<b>Regresión lineal 200-1</b>	9.740	13.801	0.007	0.997
<b>Red neuronal convolucional</b>	11.541	15.920	0.008	0.997
<b>Red neuronal LSTM</b>	14.291	20.842	0.009	0.994
<b>Regresión lineal 30-1</b>	14.752	21.226	0.010	0.994
<b>Random forest</b>	22.802	39.328	0.014	0.979

En la Tabla 15 se muestran los modelos ordenados descendenteamente por el error medio absoluto utilizando el tercer conjunto de datos. Donde igualmente se observa que el mejor modelo es la Regresión lineal.

Tabla 15. Modelos ordenados utilizados en el tercer conjunto de datos. Fuente: elaboración propia.

Modelo	MAE	RMSE	MAPE	R2
<b>Regresión lineal</b>	9.135	13.152	0.006	0.998
<b>Red neuronal BLSTM</b>	9.603	13.736	0.007	0.997
<b>Regresión lineal 200-1</b>	9.603	13.853	0.007	0.997
<b>Red neuronal densa</b>	10.078	14.267	0.007	0.997
<b>Red neuronal LSTM</b>	12.164	16.225	0.008	0.996
<b>Red neuronal convolucional</b>	12.966	18.315	0.009	0.995
<b>Regresión lineal 30-1</b>	15.034	23.798	0.010	0.992
<b>Random forest</b>	17.158	29.248	0.011	0.988

Los modelos ordenados descendenteamente por el error medio absoluto utilizando el cuarto conjunto de datos, se muestran en la Tabla 16. De igual forma se observa que el mejor modelo es la Regresión lineal.

Tabla 16. Modelos utilizando el cuarto conjunto de datos. Fuente: elaboración propia.

Modelo	MAE	RMSE	MAPE	R2
<b>Regresión lineal</b>	9.137	13.168	0.006	0.998

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>Red neuronal convolucional</b>	9.495	13.552	0.007	0.997
<b>Red neuronal densa</b>	9.507	13.501	0.007	0.998
<b>Red neuronal BLSTM</b>	9.541	13.509	0.007	0.998
<b>Regresión lineal 200-1</b>	9.672	13.808	0.007	0.997
<b>Regresión lineal 30-1</b>	14.111	21.402	0.010	0.994
<b>Random forest</b>	18.052	31.164	0.011	0.987
<b>Red neuronal LSTM</b>	21.667	34.808	0.014	0.983

En la Tabla 17 se muestran los modelos ordenados descendenteamente por el error medio absoluto utilizando el quinto conjunto de datos. La Regresión lineal es el mejor modelo otra vez.

Tabla 17. Modelos ordenados utilizados en el quinto conjunto de datos. Fuente: elaboración propia.

Modelo	MAE	RMSE	MAPE	R2
<b>Regresión lineal</b>	9.149	13.129	0.006	0.998
<b>Red neuronal BLSTM</b>	9.522	13.453	0.007	0.998
<b>Red neuronal densa</b>	9.707	13.542	0.007	0.997
<b>Red neuronal convolucional</b>	10.034	14.300	0.007	0.997
<b>Regresión lineal 200-1</b>	10.455	15.271	0.007	0.997
<b>Random forest</b>	17.387	31.078	0.011	0.987
<b>Red neuronal LSTM</b>	21.463	33.895	0.014	0.984
<b>Regresión lineal 30-1</b>	78.341	168.417	0.054	0.612

#### 5.4 Comparación de los resultados de los modelos

En esta sección se presentan los 10 mejores resultados, obtenidos para cada una de las series de precios estudiadas entre todos los modelos y los diferentes conjuntos de variables utilizadas como regresores. De los 42 modelos analizados para la predicción de precios del Oro en la Tabla 18 se muestran los 10 mejores resultados, ordenados por el error medio absoluto de forma descendente, la totalidad de los 42 modelos se muestran en la Tabla 1 del Anexo 1.

Tabla 18. Resultados de los modelos ordenados descendenteamente por MAE para el Oro. Fuente: elaboración propia.

MODELO	MAE	RMSE	MAPE	R2
<b>Gold_1er_conjunto_datos_mañana igual que ayer</b>	9.128	13.119	0.006	0.998
<b>Gold_3er_conjunto_datos_Regresión lineal</b>	9.135	13.152	0.006	0.998
<b>Gold_4to_conjunto_datos_Regresión lineal</b>	9.137	13.168	0.006	0.998
<b>Gold_2do_conjunto_datos_Regresión lineal</b>	9.141	13.112	0.006	0.998
<b>Gold_5to_conjunto_datos_Regresión lineal</b>	9.149	13.129	0.006	0.998
<b>Gold_1er_conjunto_datos_Regresión lineal</b>	9.150	13.148	0.006	0.998
<b>Gold_1er_conjunto_datos_Red neuronal densa</b>	9.397	13.318	0.006	0.998
<b>Gold_4to_conjunto_datos_Red neuronal convolucional</b>	9.495	13.552	0.007	0.997

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>Gold_4to_conjunto_datos_Red neuronal densa</b>	9.507	13.501	0.007	0.998
<b>Gold_5to_conjunto_datos_Red neuronal BLSTM</b>	9.522	13.453	0.007	0.998

Analizando los resultados obtenidos sobre la serie temporal de precios del Oro. Se pude decir que los modelos predicen bien si se toma en cuenta la métrica de R<sup>2</sup>, se comprueba que explican la variable al 99% y el error en términos de MAPE es menor a un 1%. Sin embargo, el modelo básico que predice para el próximo día es el mismo valor que el día anterior, fue el que mejor resultado obtuvo entre los 42 modelos analizados.

En la predicción de precios de la compañía Boing Company, observados en la Tabla 19 se observan los 10 mejores modelos de los 42, ordenados por el error medio absoluto de forma descendente. La totalidad de los 42 modelos se muestran en la Tabla 2 del Anexo 1.

Tabla 19. Resultados de los modelos ordenados descendente por MAE para la compañía BA. Fuente: elaboración propia.

MODELO	MAE	RMSE	MAPE	R2
<b>BA_4to_conjunto_datos_Regresión lineal</b>	3.419	5.408	0.016	0.996
<b>BA_3er_conjunto_datos_Regresión lineal</b>	3.422	5.440	0.016	0.996
<b>BA_2do_conjunto_datos_Regresión lineal</b>	3.422	5.432	0.016	0.996
<b>BA_1er_conjunto_datos_mañana igual que ayer</b>	3.426	5.430	0.016	0.996
<b>BA_1er_conjunto_datos_Regresión lineal</b>	3.439	5.450	0.016	0.996
<b>BA_5to_conjunto_datos_Regresión lineal</b>	3.441	5.449	0.016	0.996
<b>BA_4to_conjunto_datos_Red neuronal densa</b>	3.477	5.518	0.016	0.996
<b>BA_2do_conjunto_datos_Red neuronal densa</b>	3.483	5.565	0.016	0.996
<b>BA_4to_conjunto_datos_Red neuronal convolucional</b>	3.572	5.534	0.016	0.996
<b>BA_4to_conjunto_datos_Regresión lineal 200-1</b>	3.616	5.672	0.017	0.996

De los 42 modelos analizados para la predicción de precios de la compañía International Business Machines, en la Tabla 20 se muestran los 10 mejores resultados ordenados por el error medio absoluto de forma descendente, la totalidad de los 42 modelos se muestran en la Tabla 3 del Anexo 1.

Tabla 20. Resultados de los modelos de predicción ordenados descendente por MAE para la compañía IBM. Fuente: elaboración propia.

MODELO	MAE	RMSE	MAPE	R2
<b>IBM_3er_conjunto_datos_Regresión lineal</b>	1.366	2.016	0.010	0.984
<b>IBM_1er_conjunto_datos_mañana igual que ayer</b>	1.366	2.023	0.010	0.984
<b>IBM_2do_conjunto_datos_Regresión lineal</b>	1.366	2.023	0.010	0.984
<b>IBM_4to_conjunto_datos_Regresión lineal</b>	1.369	2.024	0.010	0.984
<b>IBM_5to_conjunto_datos_Regresión lineal</b>	1.370	2.021	0.010	0.984
<b>IBM_2do_conjunto_datos_Red neuronal densa</b>	1.374	2.028	0.010	0.984

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>IBM_1er_conjunto_datos_Regresión lineal</b>	1.378	2.026	0.010	0.984
<b>IBM_2do_conjunto_datos_Red neuronal BLSTM</b>	1.379	2.029	0.010	0.984
<b>IBM_4to_conjunto_datos_Red neuronal BLSTM</b>	1.386	2.040	0.010	0.984
<b>IBM_2do_conjunto_datos_Red neuronal convolucional</b>	1.389	2.053	0.010	0.984

De los 42 modelos analizados para la predicción de precios de la compañía Walt Disney Company, en la Tabla 21 se muestran los 10 mejores resultados ordenados por el error medio absoluto de forma descendente, la totalidad de los 42 modelos se muestran en la Tabla 4 del Anexo 1.

Tabla 21. Resultados de los modelos de predicción ordenados descendente por MAE para la compañía DIS. Fuente: elaboración propia.

MODELO	MAE	RMSE	MAPE	R2
<b>DIS_3er_conjunto_datos_Regresión lineal</b>	1.314	2.085	0.011	0.994
<b>DIS_2do_conjunto_datos_Regresión lineal</b>	1.314	2.082	0.011	0.994
<b>DIS_1er_conjunto_datos_mañana igual que ayer</b>	1.315	2.083	0.011	0.994
<b>DIS_4to_conjunto_datos_Regresión lineal</b>	1.315	2.084	0.011	0.994
<b>DIS_5to_conjunto_datos_Regresión lineal</b>	1.316	2.089	0.011	0.994
<b>DIS_1er_conjunto_datos_Regresión lineal</b>	1.317	2.083	0.011	0.994
<b>DIS_3er_conjunto_datos_Red neuronal densa</b>	1.332	2.100	0.011	0.994
<b>DIS_4to_conjunto_datos_Red neuronal densa</b>	1.333	2.093	0.011	0.994
<b>DIS_2do_conjunto_datos_Red neuronal densa</b>	1.356	2.126	0.011	0.994
<b>DIS_4to_conjunto_datos_Red neuronal BLSTM</b>	1.390	2.153	0.011	0.994

De los 42 modelos analizados para la predicción de precios de la compañía General Electric Company, en la Tabla 22 se muestran los 10 mejores resultados ordenados por el error medio absoluto de forma descendente, la totalidad de los 42 modelos se muestran en la Tabla 5 del Anexo 1.

Tabla 22. Resultados de los modelos de predicción ordenados descendente por MAE para la compañía GE. Fuente: elaboración propia.

MODELO	MAE	RMSE	MAPE	R2
<b>GE_1er_conjunto_datos_mañana igual que ayer</b>	1.674	2.336	0.015	0.999
<b>GE_5to_conjunto_datos_Regresión lineal</b>	1.679	2.331	0.015	0.999
<b>GE_3er_conjunto_datos_Regresión lineal</b>	1.687	2.349	0.015	0.999
<b>GE_1er_conjunto_datos_Regresión lineal</b>	1.687	2.344	0.015	0.999
<b>GE_2do_conjunto_datos_Regresión lineal</b>	1.687	2.351	0.015	0.999
<b>GE_4to_conjunto_datos_Regresión lineal</b>	1.701	2.368	0.015	0.999
<b>GE_4to_conjunto_datos_Red neuronal densa</b>	1.707	2.368	0.015	0.999
<b>GE_4to_conjunto_datos_Red neuronal convolucional</b>	1.717	2.398	0.015	0.999
<b>GE_3er_conjunto_datos_Red neuronal densa</b>	1.726	2.391	0.015	0.999
<b>GE_2do_conjunto_datos_Red neuronal densa</b>	1.726	2.404	0.015	0.999

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

## 5.5 Análisis y discusión

Una vez evaluado todos los métodos de predicción sobre cinco activos diferentes, los resultados demuestran que, aun cuando las predicciones están cercanas a los precios reales en términos porcentuales y respecto al coeficiente de determinación donde se tiene un valor cercano al óptimo. Es importante medir, no solo el rendimiento basado en métricas de errores clásicas, sino también en comparación con modelos básicos.

Estos resultados demuestran cómo, modelos de regresión lineal y modelos con un enfoque ingenuo hacen buenas aproximaciones del valor real. Se observa que, es necesario en estos casos donde se predicen precios para una serie económica valorar más las métricas de rendimiento basados en porcentajes que en valores absolutos. En el caso de la serie temporal del Oro se tiene un error medio para el mejor modelo de 9.128, y en términos de error medio porcentual es de 0.006. Mientras que en la serie de Disney el mejor error promedio absoluto es 1.314 y el porcentual es de 0.011.

## Conclusiones

En la investigación se han elaborado métodos de predicción para pronosticar el precio del Oro y los precios de acciones de cuatro empresas que cotizan en la Bolsa de Nueva York (NYSE, del inglés), que es el mayor mercado de valores del mundo en volumen monetario. Se crearon métodos sencillos, como se planteó en los objetivos de investigación, que sirvieran de base para la comparación de los modelos de aprendizaje automático. Se generaron modelos basados en Bosques Aleatorios, redes neuronales de tipo feedforward, redes convolucionales, redes recurrentes de tipo LSTM y redes neuronales recurrentes de tipo LSTM bidireccionales (BLSTM), cumpliendo así otro de los objetivos planteados. Por último, se evaluaron los métodos, utilizando métricas de rendimiento como: la raíz del error cuadrático medio, el error porcentual medio absoluto, error absoluto medio y  $R^2$ .

Se observó que la predicción de precios en series financieras, utilizando un enfoque ingenuo y métodos clásicos de regresión lineal, pueden ser métodos efectivos en una tarea extremadamente difícil como la predicción de precios de acciones en temporalidad diaria. Debido a que diferentes sucesos pueden producir fuertes cambios no esperados que afecten las cotizaciones de las empresas. Los resultados de esta investigación demuestran la complejidad de hacer predicciones en un espacio temporal de 24 horas, por lo que se considera que en próximas investigaciones se pudieran:

- Utilizar intervalos de tiempo pequeños desde un minuto hasta los 15 minutos con un enfoque de escalpar, donde un profesional de los mercados trata de operar a muy corto plazo, intra-día. Con el objetivo de estar menos expuesto en cuanto al tiempo que transcurre de una predicción a la siguiente del modelo. Estos datos se pueden obtener desde la plataforma de TradingView. Incluso sin poseer una cuenta de pago, se tiene acceso a cuatro mil datos, con cuentas de pagos se tiene acceso a datos con temporalidades en segundos y hasta 20 mil registros.
- Agregar otras variables como por ejemplo series de precios de acciones del mismo sector que tengan correlación.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

## Referencias

- Allen, F., & Karjalainen, R. J. J. o. f. E. (1999). Using genetic algorithms to find technical trading rules. *51*(2), 245-271.
- Althelaya, K. A., El-Alfy, E.-S. M., & Mohammed, S. (2018). *Evaluation of bidirectional LSTM for short-and long-term stock market prediction*. Paper presented at the 2018 9th international conference on information and communication systems (ICICS).
- Bengio, Y., Simard, P., & Frasconi, P. J. I. t. o. n. n. (1994). Learning long-term dependencies with gradient descent is difficult. *5*(2), 157-166.
- Biondo, A. E., Pluchino, A., Rapisarda, A., & Helbing, D. J. P. o. (2013). Are random trading strategies more successful than technical ones? , *8*(7), e68344.
- Bollinger, J. J. S., & Commodities. (1992). Using bollinger bands. *10*(2), 47-51.
- Booth, A., Gerding, E., & McGroarty, F. J. E. S. w. A. (2014). Automated trading with performance weighted random forests and seasonality. *41*(8), 3651-3661.
- Breiman, L. J. M. I. (2001). Random forests. *45*(1), 5-32.
- Bulmer, M. G. (1979). *Principles of statistics*: Courier Corporation.
- Carter, J. F. (2012). *Mastering the trade: Proven techniques for profiting from intraday and swing trading setups*: McGraw-Hill.
- Chong, T. T.-L., & Ng, W.-K. J. A. E. L. (2008). Technical analysis and the London stock exchange: testing the MACD and RSI rules using the FT30. *15*(14), 1111-1114.
- De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *192*, 38-48.
- Ding, G., & Qin, L. (2020). Study on the prediction of stock price based on the associated network model of LSTM. *11*(6), 1307-1317.
- Fama, E. F. J. T. j. o. F. (1970). Efficient capital markets: A review of theory and empirical work. *25*(2), 383-417.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. J. C. o. t. A. (1996). The KDD process for extracting useful knowledge from volumes of data. *39*(11), 27-34.
- Fidelity. Slow Stochastic. Retrieved from <https://www.fidelity.com/learning-center/trading-investing/technical-analysis/technical-indicator-guide/slow-stochastic>
- Gers, F. A., Schmidhuber, J., & Cummins, F. J. N. c. (2000). Learning to forget: Continual prediction with LSTM. *12*(10), 2451-2471.
- Glantz, M., & Kissell, R. L. (2013). *Multi-asset risk modeling: techniques for a global economy in an electronic and algorithmic trading era*: Academic Press.
- Gonzalez, L. (2018). Evaluando el error en los modelos de regresión. Retrieved from <https://aprendeia.com/evaluando-el-error-en-los-modelos-de-regresion/>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*: MIT press.
- Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*: Psychology Press.
- Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: A field guide to dynamical recurrent neural networks. IEEE Press In.
- Hochreiter, S., & Schmidhuber, J. J. N. c. (1997). Long short-term memory. *9*(8), 1735-1780.
- Kim, K.-j. J. N. (2003). Financial time series forecasting using support vector machines. *55*(1-2), 307-319.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. J. C. o. t. A. (2017). Imagenet classification with deep convolutional neural networks. *60*(6), 84-90.
- LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *3361*(10), 1995.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

- Lee, M.-C., Liao, J.-S., Yeh, S.-C., & Chang, J.-W. (2020). *Forecasting the Short-term Price Trend of Taiwan Stocks with Deep Neural Network*. Paper presented at the Proceedings of the 2020 11th International Conference on E-Education, E-Business, E-Management, and E-Learning.
- Livieris, I. E., Pintelas, E., & Pintelas, P. (2020). A CNN–LSTM model for gold price time-series forecasting. *32*(23), 17351-17360.
- Lo, A. W., & MacKinlay, A. C. (2011). A non-random walk down Wall Street. In *A Non-Random Walk Down Wall Street*: Princeton University Press.
- Malkiel, B. G. (1999). *A random walk down Wall Street: including a life-cycle guide to personal investing*: WW Norton & Company.
- Marvin, M., & Seymour, A. P. J. C., MA: MIT Press. (1969). *Perceptrons*. 6, 318-362.
- Mauricio, J. A. J. U. C. d. M. (2007). Análisis de series temporales.
- McCulloch, W. S., & Pitts, W. J. T. b. o. m. b. (1943). A logical calculus of the ideas immanent in nervous activity. *5*(4), 115-133.
- McNally, S., Roche, J., & Caton, S. (2018). *Predicting the price of bitcoin using machine learning*. Paper presented at the 2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP).
- Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. (2021). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. *Journal of Business & Economic Statistics*, *39*(1), 98-119. doi:10.1080/07350015.2019.1637745
- Mehtab, S., Sen, J., & Dutta, A. (2020). *Stock price prediction using machine learning and LSTM-based deep learning models*. Paper presented at the Symposium on Machine Learning and Metaheuristics Algorithms, and Applications.
- Melo, B. J. U. E. d. C. (2012). Considerações cognitivas nas técnicas de previsão no mercado financeiro.
- Nelson, D. M., Pereira, A. C., & De Oliveira, R. A. (2017). *Stock market's price movement prediction with LSTM neural networks*. Paper presented at the 2017 International joint conference on neural networks (IJCNN).
- Olah, C. (2015). Understanding Lstm Networks. Retrieved from <http://colah.github.io/posts/2015-08-Understanding-LSTMs>
- Press, G. J. F., March. (2016). Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. *23*, 15.
- Rawat, W., & Wang, Z. J. N. c. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *29*(9), 2352-2449.
- Rodrigo, J. A. Análisis de normalidad con Python. Retrieved from <https://www.cienciadedatos.net/documentos/pystats06-analisis-normalidad-python.html>
- Rodrigo, J. A. (2016). Introducción a la Regresión Lineal Múltiple. Retrieved from [https://www.cienciadedatos.net/documentos/25\\_regresion\\_lineal\\_multiple](https://www.cienciadedatos.net/documentos/25_regresion_lineal_multiple)
- Roondiwala, M., Patel, H., & Varma, S. (2017). Predicting stock prices using LSTM. *International Journal of Science Research*, *6*(4), 1754-1756.
- Rosenblatt, F. J. P. r. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *65*(6), 386.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. J. n. (1986). Learning representations by back-propagating errors. *323*(6088), 533-536.
- Schuster, M., & Paliwal, K. K. J. I. t. o. S. P. (1997). Bidirectional recurrent neural networks. *45*(11), 2673-2681.
- Shumway, R. H., Stoffer, D. S., & Stoffer, D. S. (2000). *Time series analysis and its applications* (Vol. 3): Springer.
- Silva, E., Castilho, D., Pereira, A., & Brandao, H. (2014). *A neural network based approach to support the market making strategies in high-frequency trading*. Paper presented at the 2014 International Joint Conference on Neural Networks (IJCNN).

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

- Sunny, M. A. I., Maswood, M. M. S., & Alharbi, A. G. (2020). *Deep learning-based stock price prediction using LSTM and bi-directional LSTM model*. Paper presented at the 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES).
- Valverde Nieto, E. (2021). Modelización de los precios de las acciones de Apple, Microsoft, Amazon y Google utilizando redes neuronales LSTM.
- Werbos, P. J. J. P. o. t. I. (1990). Backpropagation through time: what it does and how to do it. 78(10), 1550-1560.
- Wigglesworth, R., & Flood, C. (2018). BlackRock bulks up research into artificial intelligence. *Financial Times*.
- Yeh, C.-C., Chi, D.-J., & Lin, Y.-R. J. I. S. (2014). Going-concern prediction using hybrid random forests and rough set approach. 254, 98-110.
- Yrigoyen, C. C. (2003). *Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales*: Dirección General de Economía y Planificación.

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

## Anexo 1.

Modelos empleados en la predicción del precio del Oro ordenados por error medio absoluto descendente.

Tabla 1.

MODELO	MAE	RMSE	MAPE	R2
Gold_1er_conjunto_datos_mañana igual que ayer	9.128	13.119	0.006	0.998
Gold_3er_conjunto_datos_Regresión lineal	9.135	13.152	0.006	0.998
Gold_4to_conjunto_datos_Regresión lineal	9.137	13.168	0.006	0.998
Gold_2do_conjunto_datos_Regresión lineal	9.141	13.112	0.006	0.998
Gold_5to_conjunto_datos_Regresión lineal	9.149	13.129	0.006	0.998
Gold_1er_conjunto_datos_Regresión lineal	9.150	13.148	0.006	0.998
Gold_1er_conjunto_datos_Red neuronal densa	9.397	13.318	0.006	0.998
Gold_4to_conjunto_datos_Red neuronal convolucional	9.495	13.552	0.007	0.997
Gold_4to_conjunto_datos_Red neuronal densa	9.507	13.501	0.007	0.998
Gold_5to_conjunto_datos_Red neuronal BLSTM	9.522	13.453	0.007	0.998
Gold_4to_conjunto_datos_Red neuronal BLSTM	9.541	13.509	0.007	0.998
Gold_2do_conjunto_datos_Red neuronal densa	9.599	13.573	0.007	0.997
Gold_3er_conjunto_datos_Regresión lineal 200-1	9.603	13.853	0.007	0.997
Gold_3er_conjunto_datos_Red neuronal BLSTM	9.603	13.736	0.007	0.997
Gold_4to_conjunto_datos_Regresión lineal 200-1	9.672	13.808	0.007	0.997
Gold_1er_conjunto_datos_Red neuronal LSTM	9.683	13.602	0.007	0.997
Gold_2do_conjunto_datos_Red neuronal BLSTM	9.688	13.614	0.007	0.997
Gold_5to_conjunto_datos_Red neuronal densa	9.707	13.542	0.007	0.997
Gold_2do_conjunto_datos_Regresión lineal 200-1	9.740	13.801	0.007	0.997
Gold_1er_conjunto_datos_Red neuronal BLSTM	9.773	13.708	0.007	0.997
Gold_5to_conjunto_datos_Red neuronal convolucional	10.034	14.300	0.007	0.997
Gold_3er_conjunto_datos_Red neuronal densa	10.078	14.267	0.007	0.997
Gold_5to_conjunto_datos_Regresión lineal 200-1	10.455	15.271	0.007	0.997
Gold_1er_conjunto_datos_Red neuronal convolucional	11.129	15.465	0.008	0.997
Gold_2do_conjunto_datos_Red neuronal convolucional	11.541	15.920	0.008	0.997
Gold_1er_conjunto_datos_Regresión lineal 200-1	12.020	20.625	0.008	0.994
Gold_3er_conjunto_datos_Red neuronal LSTM	12.164	16.225	0.008	0.996
Gold_3er_conjunto_datos_Red neuronal convolucional	12.966	18.315	0.009	0.995
Gold_1er_conjunto_datos_Ayer más variación	13.483	18.354	0.009	0.995
Gold_4to_conjunto_datos_Regresión lineal 30-1	14.111	21.402	0.010	0.994
Gold_2do_conjunto_datos_Red neuronal LSTM	14.291	20.842	0.009	0.994
Gold_2do_conjunto_datos_Regresión lineal 30-1	14.752	21.226	0.010	0.994
Gold_3er_conjunto_datos_Regresión lineal 30-1	15.034	23.798	0.010	0.992
Gold_1er_conjunto_datos_Random Forest	16.878	29.408	0.011	0.988

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>Gold_3er_conjunto_datos_Random Forest</b>	<b>17.158</b>	<b>29.248</b>	<b>0.011</b>	<b>0.988</b>
<b>Gold_5to_conjunto_datos_Random Forest</b>	<b>17.387</b>	<b>31.078</b>	<b>0.011</b>	<b>0.987</b>
<b>Gold_4to_conjunto_datos_Random Forest</b>	<b>18.052</b>	<b>31.164</b>	<b>0.011</b>	<b>0.987</b>
<b>Gold_5to_conjunto_datos_Red neuronal LSTM</b>	<b>21.463</b>	<b>33.895</b>	<b>0.014</b>	<b>0.984</b>
<b>Gold_4to_conjunto_datos_Red neuronal LSTM</b>	<b>21.667</b>	<b>34.808</b>	<b>0.014</b>	<b>0.983</b>
<b>Gold_2do_conjunto_datos_Random Forest</b>	<b>22.802</b>	<b>39.328</b>	<b>0.014</b>	<b>0.979</b>
<b>Gold_1er_conjunto_datos_Regresión lineal 30-1</b>	<b>48.716</b>	<b>85.591</b>	<b>0.033</b>	<b>0.900</b>
<b>Gold_5to_conjunto_datos_Regresión lineal 30-1</b>	<b>78.341</b>	<b>168.417</b>	<b>0.054</b>	<b>0.612</b>

Modelos empleados en la predicción del precio de las acciones de la compañía IBM ordenados por error medio absoluto descendenteamente.

Tabla 2.

MODELO	MAE	RMSE	MAPE	R2
<b>IBM_3er_conjunto_datos_Regresión lineal</b>	<b>1.366</b>	<b>2.016</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_1er_conjunto_datos_mañana igual que ayer</b>	<b>1.366</b>	<b>2.023</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_2do_conjunto_datos_Regresión lineal</b>	<b>1.366</b>	<b>2.023</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_4to_conjunto_datos_Regresión lineal</b>	<b>1.369</b>	<b>2.024</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_5to_conjunto_datos_Regresión lineal</b>	<b>1.370</b>	<b>2.021</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_2do_conjunto_datos_Red neuronal densa</b>	<b>1.374</b>	<b>2.028</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_1er_conjunto_datos_Regresión lineal</b>	<b>1.378</b>	<b>2.026</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_2do_conjunto_datos_Red neuronal BLSTM</b>	<b>1.379</b>	<b>2.029</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_4to_conjunto_datos_Red neuronal BLSTM</b>	<b>1.386</b>	<b>2.040</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_2do_conjunto_datos_Red neuronal convolucional</b>	<b>1.389</b>	<b>2.053</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_4to_conjunto_datos_Red neuronal densa</b>	<b>1.390</b>	<b>2.039</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_2do_conjunto_datos_Red neuronal LSTM</b>	<b>1.392</b>	<b>2.046</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_5to_conjunto_datos_Red neuronal densa</b>	<b>1.398</b>	<b>2.056</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_5to_conjunto_datos_Red neuronal BLSTM</b>	<b>1.399</b>	<b>2.051</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_4to_conjunto_datos_Red neuronal LSTM</b>	<b>1.404</b>	<b>2.061</b>	<b>0.010</b>	<b>0.984</b>
<b>IBM_1er_conjunto_datos_Red neuronal densa</b>	<b>1.410</b>	<b>2.066</b>	<b>0.011</b>	<b>0.984</b>
<b>IBM_1er_conjunto_datos_Red neuronal LSTM</b>	<b>1.417</b>	<b>2.048</b>	<b>0.011</b>	<b>0.984</b>
<b>IBM_1er_conjunto_datos_Red neuronal BLSTM</b>	<b>1.425</b>	<b>2.070</b>	<b>0.011</b>	<b>0.983</b>
<b>IBM_2do_conjunto_datos_Regresión lineal 200-1</b>	<b>1.438</b>	<b>2.113</b>	<b>0.011</b>	<b>0.983</b>
<b>IBM_4to_conjunto_datos_Red neuronal convolucional</b>	<b>1.447</b>	<b>2.087</b>	<b>0.011</b>	<b>0.983</b>
<b>IBM_3er_conjunto_datos_Regresión lineal 200-1</b>	<b>1.447</b>	<b>2.104</b>	<b>0.011</b>	<b>0.983</b>
<b>IBM_4to_conjunto_datos_Regresión lineal 200-1</b>	<b>1.448</b>	<b>2.128</b>	<b>0.011</b>	<b>0.982</b>
<b>IBM_5to_conjunto_datos_Red neuronal convolucional</b>	<b>1.463</b>	<b>2.132</b>	<b>0.011</b>	<b>0.982</b>
<b>IBM_5to_conjunto_datos_Red neuronal LSTM</b>	<b>1.477</b>	<b>2.113</b>	<b>0.011</b>	<b>0.983</b>
<b>IBM_1er_conjunto_datos_Red neuronal convolucional</b>	<b>1.494</b>	<b>2.126</b>	<b>0.011</b>	<b>0.983</b>
<b>IBM_3er_conjunto_datos_Red neuronal convolucional</b>	<b>1.575</b>	<b>2.232</b>	<b>0.012</b>	<b>0.981</b>

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>IBM_3er_conjunto_datos_Red neuronal LSTM</b>	<b>1.579</b>	<b>2.264</b>	<b>0.012</b>	<b>0.980</b>
<b>IBM_3er_conjunto_datos_Red neuronal BLSTM</b>	<b>1.587</b>	<b>2.278</b>	<b>0.012</b>	<b>0.980</b>
<b>IBM_5to_conjunto_datos_Regresión lineal 200-1</b>	<b>1.597</b>	<b>2.371</b>	<b>0.012</b>	<b>0.978</b>
<b>IBM_1er_conjunto_datos_Regresión lineal 200-1</b>	<b>1.778</b>	<b>2.615</b>	<b>0.013</b>	<b>0.974</b>
<b>IBM_2do_conjunto_datos_Random Forest</b>	<b>1.806</b>	<b>2.499</b>	<b>0.013</b>	<b>0.976</b>
<b>IBM_3er_conjunto_datos_Red neuronal densa</b>	<b>1.868</b>	<b>2.705</b>	<b>0.014</b>	<b>0.972</b>
<b>IBM_2do_conjunto_datos_Regresión lineal 30-1</b>	<b>1.940</b>	<b>2.804</b>	<b>0.014</b>	<b>0.970</b>
<b>IBM_4to_conjunto_datos_Random Forest</b>	<b>1.951</b>	<b>2.628</b>	<b>0.014</b>	<b>0.973</b>
<b>IBM_1er_conjunto_datos_Ayer más variación</b>	<b>2.033</b>	<b>2.916</b>	<b>0.015</b>	<b>0.967</b>
<b>IBM_5to_conjunto_datos_Random Forest</b>	<b>2.196</b>	<b>3.024</b>	<b>0.016</b>	<b>0.965</b>
<b>IBM_1er_conjunto_datos_Random Forest</b>	<b>2.334</b>	<b>3.180</b>	<b>0.017</b>	<b>0.961</b>
<b>IBM_3er_conjunto_datos_Regresión lineal 30-1</b>	<b>2.493</b>	<b>3.897</b>	<b>0.019</b>	<b>0.941</b>
<b>IBM_3er_conjunto_datos_Random Forest</b>	<b>3.877</b>	<b>5.776</b>	<b>0.029</b>	<b>0.871</b>
<b>IBM_1er_conjunto_datos_Regresión lineal 30-1</b>	<b>7.307</b>	<b>12.372</b>	<b>0.055</b>	<b>0.410</b>
<b>IBM_5to_conjunto_datos_Regresión lineal 30-1</b>	<b>13.901</b>	<b>37.197</b>	<b>0.106</b>	<b>-4.334</b>

Modelos empleados en la predicción del precio de las acciones de la compañía General Electric Company ordenados por error medio absoluto descendenteamente.

Tabla 3.

MODELO	MAE	RMSE	MAPE	R2
<b>GE_1er_conjunto_datos_mañana igual que ayer</b>	<b>1.674</b>	<b>2.336</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_5to_conjunto_datos_Regresión lineal</b>	<b>1.679</b>	<b>2.331</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_3er_conjunto_datos_Regresión lineal</b>	<b>1.687</b>	<b>2.349</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_1er_conjunto_datos_Regresión lineal</b>	<b>1.687</b>	<b>2.344</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_2do_conjunto_datos_Regresión lineal</b>	<b>1.687</b>	<b>2.351</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_4to_conjunto_datos_Regresión lineal</b>	<b>1.701</b>	<b>2.368</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_4to_conjunto_datos_Red neuronal densa</b>	<b>1.707</b>	<b>2.368</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_4to_conjunto_datos_Red neuronal convolucional</b>	<b>1.717</b>	<b>2.398</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_3er_conjunto_datos_Red neuronal densa</b>	<b>1.726</b>	<b>2.391</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_2do_conjunto_datos_Red neuronal densa</b>	<b>1.726</b>	<b>2.404</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_1er_conjunto_datos_Red neuronal LSTM</b>	<b>1.728</b>	<b>2.395</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_4to_conjunto_datos_Red neuronal BLSTM</b>	<b>1.735</b>	<b>2.393</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_5to_conjunto_datos_Red neuronal BLSTM</b>	<b>1.736</b>	<b>2.386</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_4to_conjunto_datos_Red neuronal LSTM</b>	<b>1.738</b>	<b>2.401</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_2do_conjunto_datos_Red neuronal convolucional</b>	<b>1.741</b>	<b>2.392</b>	<b>0.015</b>	<b>0.999</b>
<b>GE_1er_conjunto_datos_Red neuronal BLSTM</b>	<b>1.743</b>	<b>2.404</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_2do_conjunto_datos_Regresión lineal 200-1</b>	<b>1.750</b>	<b>2.442</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_5to_conjunto_datos_Red neuronal densa</b>	<b>1.751</b>	<b>2.405</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_4to_conjunto_datos_Regresión lineal 200-1</b>	<b>1.769</b>	<b>2.480</b>	<b>0.016</b>	<b>0.999</b>

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>GE_3er_conjunto_datos_Regresión lineal 200-1</b>	<b>1.781</b>	<b>2.470</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_3er_conjunto_datos_Red neuronal BLSTM</b>	<b>1.783</b>	<b>2.455</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_2do_conjunto_datos_Red neuronal LSTM</b>	<b>1.787</b>	<b>2.467</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_5to_conjunto_datos_Red neuronal LSTM</b>	<b>1.803</b>	<b>2.467</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_1er_conjunto_datos_Red neuronal densa</b>	<b>1.805</b>	<b>2.475</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_2do_conjunto_datos_Red neuronal BLSTM</b>	<b>1.807</b>	<b>2.490</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_5to_conjunto_datos_Red neuronal convolucional</b>	<b>1.807</b>	<b>2.476</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_3er_conjunto_datos_Red neuronal LSTM</b>	<b>1.826</b>	<b>2.489</b>	<b>0.016</b>	<b>0.999</b>
<b>GE_1er_conjunto_datos_Red neuronal convolucional</b>	<b>1.929</b>	<b>2.649</b>	<b>0.018</b>	<b>0.998</b>
<b>GE_5to_conjunto_datos_Random Forest</b>	<b>1.948</b>	<b>2.636</b>	<b>0.017</b>	<b>0.998</b>
<b>GE_1er_conjunto_datos_Random Forest</b>	<b>1.967</b>	<b>2.606</b>	<b>0.018</b>	<b>0.998</b>
<b>GE_5to_conjunto_datos_Regresión lineal 200-1</b>	<b>1.975</b>	<b>2.792</b>	<b>0.018</b>	<b>0.998</b>
<b>GE_2do_conjunto_datos_Random Forest</b>	<b>1.978</b>	<b>2.679</b>	<b>0.017</b>	<b>0.998</b>
<b>GE_4to_conjunto_datos_Random Forest</b>	<b>2.013</b>	<b>2.697</b>	<b>0.018</b>	<b>0.998</b>
<b>GE_3er_conjunto_datos_Random Forest</b>	<b>2.090</b>	<b>2.776</b>	<b>0.018</b>	<b>0.998</b>
<b>GE_3er_conjunto_datos_Red neuronal convolucional</b>	<b>2.211</b>	<b>2.955</b>	<b>0.018</b>	<b>0.998</b>
<b>GE_1er_conjunto_datos_Regresión lineal 200-1</b>	<b>2.218</b>	<b>3.160</b>	<b>0.020</b>	<b>0.998</b>
<b>GE_1er_conjunto_datos_Ayer más variación</b>	<b>2.381</b>	<b>3.269</b>	<b>0.021</b>	<b>0.998</b>
<b>GE_2do_conjunto_datos_Regresión lineal 30-1</b>	<b>2.772</b>	<b>3.901</b>	<b>0.024</b>	<b>0.996</b>
<b>GE_3er_conjunto_datos_Regresión lineal 30-1</b>	<b>2.910</b>	<b>4.299</b>	<b>0.026</b>	<b>0.996</b>
<b>GE_1er_conjunto_datos_Regresión lineal 30-1</b>	<b>9.356</b>	<b>22.213</b>	<b>0.085</b>	<b>0.886</b>
<b>GE_5to_conjunto_datos_Regresión lineal 30-1</b>	<b>15.114</b>	<b>39.374</b>	<b>0.139</b>	<b>0.640</b>

Modelos empleados en la predicción del precio de las acciones de la compañía Boing Company ordenados por error medio absoluto descendenteamente.

Tabla 4.

MODELO	MAE	RMSE	MAPE	R2
<b>BA_4to_conjunto_datos_Regresión lineal</b>	<b>3.419</b>	<b>5.408</b>	<b>0.016</b>	<b>0.996</b>
<b>BA_3er_conjunto_datos_Regresión lineal</b>	<b>3.422</b>	<b>5.440</b>	<b>0.016</b>	<b>0.996</b>
<b>BA_2do_conjunto_datos_Regresión lineal</b>	<b>3.422</b>	<b>5.432</b>	<b>0.016</b>	<b>0.996</b>
<b>BA_1er_conjunto_datos_mañana igual que ayer</b>	<b>3.426</b>	<b>5.430</b>	<b>0.016</b>	<b>0.996</b>
<b>BA_1er_conjunto_datos_Regresión lineal</b>	<b>3.439</b>	<b>5.450</b>	<b>0.016</b>	<b>0.996</b>
<b>BA_5to_conjunto_datos_Regresión lineal</b>	<b>3.441</b>	<b>5.449</b>	<b>0.016</b>	<b>0.996</b>
<b>BA_4to_conjunto_datos_Red neuronal densa</b>	<b>3.477</b>	<b>5.518</b>	<b>0.016</b>	<b>0.996</b>
<b>BA_2do_conjunto_datos_Red neuronal densa</b>	<b>3.483</b>	<b>5.565</b>	<b>0.016</b>	<b>0.996</b>
<b>BA_4to_conjunto_datos_Red neuronal convolucional</b>	<b>3.572</b>	<b>5.534</b>	<b>0.016</b>	<b>0.996</b>
<b>BA_4to_conjunto_datos_Regresión lineal 200-1</b>	<b>3.616</b>	<b>5.672</b>	<b>0.017</b>	<b>0.996</b>
<b>BA_2do_conjunto_datos_Regresión lineal 200-1</b>	<b>3.656</b>	<b>5.775</b>	<b>0.017</b>	<b>0.996</b>
<b>BA_3er_conjunto_datos_Red neuronal densa</b>	<b>3.683</b>	<b>5.816</b>	<b>0.017</b>	<b>0.996</b>

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>BA_3er_conjunto_datos_Regresión lineal 200-1</b>	<b>3.701</b>	<b>5.874</b>	<b>0.017</b>	<b>0.996</b>
<b>BA_4to_conjunto_datos_Red neuronal BLSTM</b>	<b>3.720</b>	<b>5.784</b>	<b>0.017</b>	<b>0.996</b>
<b>BA_1er_conjunto_datos_Red neuronal densa</b>	<b>3.730</b>	<b>5.857</b>	<b>0.017</b>	<b>0.996</b>
<b>BA_2do_conjunto_datos_Red neuronal convolucional</b>	<b>3.843</b>	<b>6.073</b>	<b>0.018</b>	<b>0.995</b>
<b>BA_3er_conjunto_datos_Red neuronal BLSTM</b>	<b>3.965</b>	<b>6.213</b>	<b>0.018</b>	<b>0.995</b>
<b>BA_5to_conjunto_datos_Regresión lineal 200-1</b>	<b>4.185</b>	<b>7.696</b>	<b>0.020</b>	<b>0.993</b>
<b>BA_1er_conjunto_datos_Red neuronal convolucional</b>	<b>4.306</b>	<b>7.116</b>	<b>0.020</b>	<b>0.994</b>
<b>BA_5to_conjunto_datos_Red neuronal convolucional</b>	<b>4.473</b>	<b>6.986</b>	<b>0.020</b>	<b>0.994</b>
<b>BA_1er_conjunto_datos_Regresión lineal 200-1</b>	<b>4.631</b>	<b>8.346</b>	<b>0.022</b>	<b>0.991</b>
<b>BA_2do_conjunto_datos_Red neuronal BLSTM</b>	<b>4.677</b>	<b>6.931</b>	<b>0.021</b>	<b>0.994</b>
<b>BA_5to_conjunto_datos_Red neuronal densa</b>	<b>4.680</b>	<b>7.227</b>	<b>0.020</b>	<b>0.993</b>
<b>BA_5to_conjunto_datos_Red neuronal BLSTM</b>	<b>4.699</b>	<b>7.319</b>	<b>0.020</b>	<b>0.993</b>
<b>BA_1er_conjunto_datos_Ayer más variación</b>	<b>4.907</b>	<b>7.567</b>	<b>0.022</b>	<b>0.993</b>
<b>BA_2do_conjunto_datos_Regresión lineal 30-1</b>	<b>4.920</b>	<b>7.836</b>	<b>0.023</b>	<b>0.992</b>
<b>BA_3er_conjunto_datos_Regresión lineal 30-1</b>	<b>5.611</b>	<b>9.295</b>	<b>0.026</b>	<b>0.989</b>
<b>BA_1er_conjunto_datos_Red neuronal LSTM</b>	<b>6.492</b>	<b>9.700</b>	<b>0.026</b>	<b>0.988</b>
<b>BA_3er_conjunto_datos_Red neuronal convolucional</b>	<b>6.660</b>	<b>10.231</b>	<b>0.028</b>	<b>0.987</b>
<b>BA_1er_conjunto_datos_Red neuronal BLSTM</b>	<b>7.646</b>	<b>11.273</b>	<b>0.029</b>	<b>0.984</b>
<b>BA_1er_conjunto_datos_Regresión lineal 30-1</b>	<b>16.166</b>	<b>27.871</b>	<b>0.076</b>	<b>0.902</b>
<b>BA_2do_conjunto_datos_Red neuronal LSTM</b>	<b>30.125</b>	<b>46.465</b>	<b>0.099</b>	<b>0.728</b>
<b>BA_5to_conjunto_datos_Regresión lineal 30-1</b>	<b>32.959</b>	<b>85.991</b>	<b>0.152</b>	<b>0.069</b>
<b>BA_3er_conjunto_datos_Red neuronal LSTM</b>	<b>43.840</b>	<b>64.101</b>	<b>0.150</b>	<b>0.482</b>
<b>BA_4to_conjunto_datos_Red neuronal LSTM</b>	<b>72.529</b>	<b>106.955</b>	<b>0.240</b>	<b>-0.441</b>
<b>BA_5to_conjunto_datos_Red neuronal LSTM</b>	<b>81.066</b>	<b>115.662</b>	<b>0.275</b>	<b>-0.685</b>
<b>BA_3er_conjunto_datos_Random Forest</b>	<b>86.120</b>	<b>121.777</b>	<b>0.293</b>	<b>-0.868</b>
<b>BA_2do_conjunto_datos_Random Forest</b>	<b>86.250</b>	<b>121.891</b>	<b>0.294</b>	<b>-0.871</b>
<b>BA_4to_conjunto_datos_Random Forest</b>	<b>87.038</b>	<b>122.629</b>	<b>0.297</b>	<b>-0.894</b>
<b>BA_1er_conjunto_datos_Random Forest</b>	<b>87.102</b>	<b>122.645</b>	<b>0.298</b>	<b>-0.895</b>
<b>BA_5to_conjunto_datos_Random Forest</b>	<b>87.171</b>	<b>122.688</b>	<b>0.298</b>	<b>-0.896</b>

Modelos empleados en la predicción del precio de las acciones de la compañía Walt Disney Company ordenados por error medio absoluto descendenteamente.

Tabla 5.

MODELO	MAE	RMSE	MAPE	R2
<b>DIS_3er_conjunto_datos_Regresión lineal</b>	<b>1.314</b>	<b>2.085</b>	<b>0.011</b>	<b>0.994</b>
<b>DIS_2do_conjunto_datos_Regresión lineal</b>	<b>1.314</b>	<b>2.082</b>	<b>0.011</b>	<b>0.994</b>
<b>DIS_1er_conjunto_datos_mañana igual que ayer</b>	<b>1.315</b>	<b>2.083</b>	<b>0.011</b>	<b>0.994</b>
<b>DIS_4to_conjunto_datos_Regresión lineal</b>	<b>1.315</b>	<b>2.084</b>	<b>0.011</b>	<b>0.994</b>
<b>DIS_5to_conjunto_datos_Regresión lineal</b>	<b>1.316</b>	<b>2.089</b>	<b>0.011</b>	<b>0.994</b>

“Es fácil mentir con estadísticas. Es difícil decir la verdad sin estadísticas.” Andrejs Dunkels

<b>DIS_1er_conjunto_datos_Regresión lineal</b>	<b>1.317</b>	<b>2.083</b>	<b>0.011</b>	<b>0.994</b>
<b>DIS_3er_conjunto_datos_Red neuronal densa</b>	<b>1.332</b>	<b>2.100</b>	<b>0.011</b>	<b>0.994</b>
<b>DIS_4to_conjunto_datos_Red neuronal densa</b>	<b>1.333</b>	<b>2.093</b>	<b>0.011</b>	<b>0.994</b>
<b>DIS_2do_conjunto_datos_Red neuronal densa</b>	<b>1.356</b>	<b>2.126</b>	<b>0.011</b>	<b>0.994</b>
<b>DIS_4to_conjunto_datos_Red neuronal BLSTM</b>	<b>1.390</b>	<b>2.153</b>	<b>0.011</b>	<b>0.994</b>
<b>DIS_4to_conjunto_datos_Regresión lineal 200-1</b>	<b>1.391</b>	<b>2.169</b>	<b>0.011</b>	<b>0.993</b>
<b>DIS_2do_conjunto_datos_Regresión lineal 200-1</b>	<b>1.396</b>	<b>2.189</b>	<b>0.011</b>	<b>0.993</b>
<b>DIS_3er_conjunto_datos_Regresión lineal 200-1</b>	<b>1.415</b>	<b>2.211</b>	<b>0.012</b>	<b>0.993</b>
<b>DIS_3er_conjunto_datos_Red neuronal BLSTM</b>	<b>1.455</b>	<b>2.243</b>	<b>0.012</b>	<b>0.993</b>
<b>DIS_1er_conjunto_datos_Red neuronal densa</b>	<b>1.484</b>	<b>2.262</b>	<b>0.012</b>	<b>0.993</b>
<b>DIS_5to_conjunto_datos_Red neuronal BLSTM</b>	<b>1.535</b>	<b>2.393</b>	<b>0.012</b>	<b>0.992</b>
<b>DIS_5to_conjunto_datos_Regresión lineal 200-1</b>	<b>1.553</b>	<b>2.456</b>	<b>0.013</b>	<b>0.992</b>
<b>DIS_5to_conjunto_datos_Red neuronal densa</b>	<b>1.585</b>	<b>2.398</b>	<b>0.013</b>	<b>0.992</b>
<b>DIS_5to_conjunto_datos_Red neuronal convolucional</b>	<b>1.586</b>	<b>2.458</b>	<b>0.013</b>	<b>0.992</b>
<b>DIS_1er_conjunto_datos_Red neuronal convolucional</b>	<b>1.588</b>	<b>2.488</b>	<b>0.013</b>	<b>0.991</b>
<b>DIS_2do_conjunto_datos_Red neuronal convolucional</b>	<b>1.610</b>	<b>2.546</b>	<b>0.013</b>	<b>0.991</b>
<b>DIS_1er_conjunto_datos_Red neuronal BLSTM</b>	<b>1.661</b>	<b>2.578</b>	<b>0.013</b>	<b>0.991</b>
<b>DIS_2do_conjunto_datos_Red neuronal BLSTM</b>	<b>1.687</b>	<b>2.430</b>	<b>0.014</b>	<b>0.992</b>
<b>DIS_1er_conjunto_datos_Regresión lineal 200-1</b>	<b>1.719</b>	<b>2.659</b>	<b>0.014</b>	<b>0.990</b>
<b>DIS_4to_conjunto_datos_Red neuronal convolucional</b>	<b>1.797</b>	<b>2.658</b>	<b>0.015</b>	<b>0.990</b>
<b>DIS_1er_conjunto_datos_Red neuronal LSTM</b>	<b>1.840</b>	<b>2.874</b>	<b>0.014</b>	<b>0.988</b>
<b>DIS_1er_conjunto_datos_Ayer más variación</b>	<b>1.960</b>	<b>3.046</b>	<b>0.016</b>	<b>0.987</b>
<b>DIS_2do_conjunto_datos_Regresión lineal 30-1</b>	<b>1.988</b>	<b>3.070</b>	<b>0.016</b>	<b>0.987</b>
<b>DIS_3er_conjunto_datos_Red neuronal convolucional</b>	<b>2.199</b>	<b>3.217</b>	<b>0.017</b>	<b>0.986</b>
<b>DIS_3er_conjunto_datos_Regresión lineal 30-1</b>	<b>2.213</b>	<b>3.373</b>	<b>0.018</b>	<b>0.984</b>
<b>DIS_1er_conjunto_datos_Regresión lineal 30-1</b>	<b>6.284</b>	<b>11.755</b>	<b>0.051</b>	<b>0.807</b>
<b>DIS_2do_conjunto_datos_Red neuronal LSTM</b>	<b>7.074</b>	<b>11.366</b>	<b>0.048</b>	<b>0.820</b>
<b>DIS_3er_conjunto_datos_Red neuronal LSTM</b>	<b>8.970</b>	<b>14.540</b>	<b>0.061</b>	<b>0.705</b>
<b>DIS_5to_conjunto_datos_Regresión lineal 30-1</b>	<b>12.085</b>	<b>32.079</b>	<b>0.097</b>	<b>-0.435</b>
<b>DIS_4to_conjunto_datos_Red neuronal LSTM</b>	<b>18.123</b>	<b>27.092</b>	<b>0.126</b>	<b>-0.024</b>
<b>DIS_5to_conjunto_datos_Red neuronal LSTM</b>	<b>31.114</b>	<b>40.720</b>	<b>0.227</b>	<b>-1.312</b>
<b>DIS_4to_conjunto_datos_Random Forest</b>	<b>34.610</b>	<b>43.764</b>	<b>0.256</b>	<b>-1.671</b>
<b>DIS_2do_conjunto_datos_Random Forest</b>	<b>34.786</b>	<b>43.900</b>	<b>0.258</b>	<b>-1.687</b>
<b>DIS_5to_conjunto_datos_Random Forest</b>	<b>34.887</b>	<b>43.975</b>	<b>0.258</b>	<b>-1.696</b>
<b>DIS_1er_conjunto_datos_Random Forest</b>	<b>34.894</b>	<b>43.986</b>	<b>0.258</b>	<b>-1.697</b>
<b>DIS_3er_conjunto_datos_Random Forest</b>	<b>34.927</b>	<b>44.016</b>	<b>0.259</b>	<b>-1.701</b>